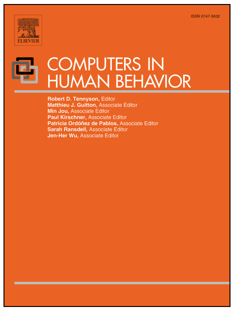
Accepted Manuscript



ISSN 0747-5632

ELSEVIER

COMPUTERS IN HUMAN BEHAVIOR

Egoistic and altruistic motivation: How to induce users' willingness to help for imperfect AI

Yeonjoo Lee, Miyeon Ha, Sujeong Kwon, Yealin Shim, Jinwoo Kim

PII: S0747-5632(19)30227-4

DOI: https://doi.org/10.1016/j.chb.2019.06.009

Reference: CHB 6049

To appear in: *Computers in Human Behavior*

Received Date: 14 January 2019

Revised Date: 11 April 2019

Accepted Date: 19 June 2019

Please cite this article as: Lee Y., Ha M., Kwon S., Shim Y. & Kim J., Egoistic and altruistic motivation: How to induce users' willingness to help for imperfect AI, *Computers in Human Behavior* (2019), doi: https://doi.org/10.1016/j.chb.2019.06.009.

Egoistic and Altruistic Motivation:
How to Induce Users' Willingness to Help for Imperfect AI

Yeonjoo Lee[1], Miyeon Ha[2], Sujeong Kwon[3], Yealin Shim[4], Jinwoo Kim[5]

HCI lab, Yonsei University
HCI Lab, 502, Business Hall, Yonsei University, Yonsei-ro, Seodaemun-gu, Seoul, 120-749, South Korea

[1] yunjoo425100@gmail.com, [2] shelling0622@gmail.com, [3] kwonsujeong96@gmail.com, [4] yealinshim@gmail.com, [5] jinwoo@yonsei.ac.kr

Corresponding Author:
Jinwoo Kim
Yonsei University
HCI Lab, 502, Business Hall, Yonsei University, Yonsei-ro, Seodaemun-gu, Seoul, 120-749, South Korea
+82 10-6307-2528
jinwoo@yonsei.ac.kr

Acknowledgements:
This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2016R1D1A1B02015987). Declarations of interest: none.

Computers in Human Behavior

Egoistic and Altruistic Motivation: How to Induce Users' Willingness to Help for Imperfect AI

Keywords

AI (Artificial Intelligence); Human in the Loop (HITL); Willingness to Help; Trust; Explainable AI; Food Image Recognition

ABSTRACT

Although artificial intelligence is a growing area of research, several problems remain. One such problem of particular importance is the low accuracy of predictions. This paper suggests that users' help is a practical approach to improve accuracy and it considers four factors that trigger users' willingness to help for an imperfect AI system. The two factors covered in Study 1 are utilitarian benefit based on egoistic motivation, and empathy based on altruistic motivation. In Study 2, utilitarian benefit is divided into explainable AI and monetary reward. The results indicate that two variables, namely empathy and monetary reward, have significant positive effects on willingness to help, and monetary reward is the strongest stimulus. In addition, explainable AI is shown to be positively associated with trust in AI. This study applies social studies of help motivation to the HCI field in order to induce users' willingness to help for an imperfect AI. The triggers of help motivation, empathy and monetary reward, can be utilized to induce the users' voluntary engagement in the loop with an imperfect AI.

1. Introduction

Artificial Intelligence (AI) has progressed rapidly with the advancement of algorithms due to the vast amount of data now available (O'Leary, 2013). However, AI is still deficient in terms of the accuracy of its predictions, preventing it from providing users with a seamless AI experience (Shoham, Perrault, Brynjolfsson, & Clark, 2017). This is a vital issue that must be resolved because when the accuracy of a system is not guaranteed, the user experience deteriorates (Dill & Rabin, 2013). Properly labeled and structured data is the fundamental source of the AI learning process (Z. Liu, Qiao, Long, & Li, 2018; O'Leary, 2013). Particularly, human-labeled data is effective in the enhancement of AI when the AI itself cannot achieve a high level of accuracy (Imran, Castillo, Lucas, Meier, & Vieweg, 2014). Humans are capable of producing high-quality data that AI lacks, including complex image recognition, speech recognition, and translation in the field (Shoham et al., 2017). Thus, the concept of human-in-the-loop (HITL) (DEFENSE, 1994) has been suggested as a promising way to acquire high-quality data for the advancement of AI performance (Fails & Olsen Jr, 2003; Holzinger, 2016; Lasecki et al., 2013; Rahwan, 2018). Recently, HITL has been used synonymously with interactive machine learning (IML); here, human-interventions, such as the correction of AI predictions, are adjusted interactively. IML improves the accuracy of AI quickly and effectively as human knowledge is instantly reflected in the algorithm (Fails & Olsen Jr, 2003). Although earlier studies have highlighted the importance of human intervention, research into how humans can be induced to participate in the loop has received scant attention. Therefore, this paper aims to investigate the factors that heighten users' willingness to help an imperfect AI system.

In order to identify factors that induce higher levels of willingness to help, we investigated two main motives for help that are studied in the field of social science: egoistic and altruistic motivation (Batson, O'Quin, Fultz, Vanderplas, & Isen, 1983). The egoistic motivation was designed to be driven by benefit, while the altruistic motivation was intended to be triggered by empathy. However, empathy is not the only aspect of altruistic motivation. Altruism can result from the prospect of economic gain (Simon, 1993), self-satisfaction, or the reduction of aversive feedback (Batson, 1987). Nevertheless, we selected empathy as the trigger for altruistic motivation because empathy is an indispensable component of altruistic behavior. Batson, Duncan, Ackerman, Buckley, and Birch (1981) implemented an experiment revealed that empathy led to the altruistic motivation to help. When participants felt high levels of empathy, they helped the same amount regardless of whether they could have easily avoided the situation or not. On the other hand, when low levels of empathy were induced, participants helped only when it was difficult to evade the situation. This demonstrates that empathy is a critical element of human beings' altruistic behavior. From the egoistic motivation perspective, a benefit was provided because it is considered to be a robust motive for human behavior (Vallerand, 1997). For altruistic motivation, empathy for the AI agent was included, which was posited to encourage users to help even without receiving direct benefits (Misselhorn, 2009). In addition, trust in the AI is hypothesized to affect users' decisions to participate in the loop given the two motivations; trust is an important factor for interhuman help and interactions, so trust in a system affects one's intention to help (Twenge, Baumeister, DeWall, Ciarocco, & Bartels, 2007).

We chose to use AI food image recognition as the application domain to test the impacts of egoistic and altruistic motivations on users' willingness to help. Despite its convenience, fully automatic food journaling via image recognition is imperfect in terms of accuracy, limiting the use of AI food journaling (C. Liu et al., 2016; Meyers et al., 2015). However, ordinary people can discern types of certain foods without intensive training. Therefore, food recognition is a domain where ordinary people can provide actual help for an imperfect AI. Therefore, we developed a food-recognizing prototype mobile application, and conducted two studies regarding the following research questions:

1. How do egoistic motivation, altruistic motivation and trust affect users' willingness to help an AI system?
2. How do egoistic motivation and altruistic motivation affect users' trust in an AI system?
3. How does trust mediate the effects of egoistic motivation and altruistic motivation on users' willingness to help an AI system?

Study 1 revealed that empathy, the trigger for altruistic motivation, has a direct positive effect on willingness to help, whereas utilitarian benefit, a type of egoistic motivation, has an indirect positive effect by means of trust. In other words, unlike empathy, the direct effect of the utilitarian benefit was not found. This is because the utilitarian benefit construct had two different sub-constructs, explainable AI and performance improvement. Although both sub-constructs were designed to act as a benefit in the form of an explanation, explainable AI is closer to the concept of an explanation, while the performance improvement is closer to the concept of a benefit. Accordingly, the need for further investigation into the separate effects of the two sub-constructs—explainable AI and performance improvement—arose. Moreover, the participants did not regard the performance improvement provided in Study 1 as being directly advantageous to them. Therefore, in Study 2, we separated the explainable AI from the utilitarian benefit and changed performance improvement into a monetary reward, which is a more powerful type of benefit. Study 2 revealed that a monetary reward positively affects willingness to help and explainable AI is positively associated with trust.

This work demonstrates that specific factors can elicit users' trust and willingness to help. This can be applied practically to the design of AI services that are still deficient in terms of accuracy. The rest of this paper is organized as follows: Section 2 reviews related previous studies, and Section 3 builds hypotheses based on Section 2. Sections 4 and 5 explain the processes, results, and limitations of Studies 1 and 2, respectively. Lastly, Section 6 discusses the overall implications and limitations of this research.

2. Background

2.1. Imperfections of AI and Food Image Recognition

AI is often reported to surpass human ability, but human knowledge is still superior to AI in various domains such as complicated image recognition, speech recognition, and translation (Shoham et al., 2017). Specifically, in the field of image recognition, which has made a remarkable progress since 2012, there are some limitations in terms of accuracy. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is the main source used to evaluate the progress of image recognition (Russakovsky et al., 2015). The limitation of this challenge is that its accuracy varies significantly between image categories. Although a decent performance was achieved overall, the performance in a number of categories fell behind. This indicates that achieving high accuracy across all categories has not been realized yet.

AI food image recognition is far behind the average prediction accuracy of ILSVRC for 'Image Classification' tasks, which is 94.6 %. Liu et al. (2016) trained two food image data sets, UEC-256 and Food-101, and achieved a maximum accuracy of 77.4 %. The latest study of food recognition argues that the accuracy of algorithms falls below 80 % when they are shown together (Hoff, Jaffurs, Enriquez, & Wilde, 2018). This is because object detection that predicts multiple objects at once has a precision level of 44.7 %. Object detection skill is required for real-life food image recognition because it is not natural to take photos of food one at a time. When people have a meal, they usually have a main dish, a side dish, and a drink. Thus, the high accuracy rate of ILSVRC is not applicable to food image recognition. Moreover, food images are not abundant in the ImageNet database. Furthermore, the diversity of food among different cultures is too great to train every single image of food globally. As a result, the accuracy of food prediction is low; to overcome this limitation, it is crucial to collect a vast amount of data regarding food. This study suggests that interactions that induce human input following incorrect predictions are a way to collect such data, and eventually to enhance the AI algorithm. Existing AI systems are not perfect, and allowing users to fix the incorrect predictions of food-recognizing AI has been indicated as a potential future solution for the development of food image processing algorithms (Hoff et al., 2018).

2.2. Human-in-the-Loop (HITL)

In a broad sense, HITL refers to models that include human interaction (DEFENSE, 1994). This type of HITL incorporates human workers in the system for Turing tests (Naor, 1996), to evaluate and edit machine translations (Callison-Burch, 2009), and to provide cognitive skills that are deficient in automatic systems (Leeper, Hsiao, Ciocarlie, Takayama, & Gossow, 2012). Although human workers can help a system, their help is limited to evaluating or supporting system performance in real-time with their cognitive skills or knowledge. Recently, HITL has been used in a narrower sense, closely related to AI and machine learning. HITL is used synonymously with IML (Interactive Machine Learning), which is defined as a machine learning system that optimizes its performance through interaction with other agents such as human beings (Holzinger, 2016). Previous studies of IML deem human intervention to be a crucial component that oversees, regulates, and optimizes AI, as shown in **Table 1**. Unlike HITL in a broad sense, HITL as IML generally allows human intervention to be utilized as training data. For example, human workers can up- or down-vote to decide whether a chatbot's answer is appropriate or not. When this human labeled data is used for training to form an automated voting system, the overall system performance can be improved substantially and human workers can gradually be excluded from the voting process (Lasecki et al., 2013). Thus, interacting with a user throughout the process can establish an AI system efficiently.

Table 1 Overview of prior studies regarding HITL

Study	Definition of HITL	Role of Human	Topic & Domain
(Fails & Olsen Jr, 2003)	Fast and focused training system facilitated by iterative human corrections for a classifier	Train, classify/view, and correct classifications made by AI	Image processing with interactive machine learning system, "Crayons"
(Lasecki et al., 2013)	Supervised machine learning approach that provides oversight, controls data quality, and provides human data	Vote, answer	Crowd-Powered Conversational Assistant (AI)
(Holzinger , 2016)	Algorithms that interact with agents, including humans, and optimize their learning behavior based on these interactions	Provide human expertise for system efficiency and accuracy	Healthcare
(Rahwan, 2018)	An automated control process where a human operator is a crucial component, who handles supervision, exception control, optimization, and maintenance	Identify misbehavior, provide an accountable entity	Society-in-the-loop
(Zanzotto, 2017)	A fairer paradigm for AI system that facilitates a transparent knowledge lifecycle and provides reward for the owner of the data	Provide human generated data and receive profit	HIT-AI (Human-in-the-loop + AI)
(Brown & Grinter,	Human-in-the-loop for real-time translation	Translate	Translation platform

2016)			
(Callison-Burch, 2009)	Human-in-the-loop for improving machine translation	Evaluate and edit machine translation	Human mediated translation edit rate
(Leeper et al., 2012)	Human-in-the-loop for providing human knowledge	Provide cognitive skills for robot and command	Robotics
(Naor, 1996)	Human-in-the-loop to tell humans and computers apart	Turing test	Turing test

For an AI with deficits, IML is a better training model than classical machine learning (CML). CML requires feature selection prior to the training process; accordingly, CML classifiers cannot be generated interactively, leading to significant dependence on the initial feature selection process. On the other hand, IML interacts with users and utilizes their corrections to eliminate classifier errors. Thus, IML develops appropriate classifiers efficiently and rapidly (Fails & Olsen Jr, 2003). If an AI algorithm with low accuracy is created using CML, adjusting the feature selection process and supplementing data, which does not always guarantee improved performance, are the only options for change. Thus, IML, which allows human intervention for corrections throughout the process, is considered a more effective option for a defective AI. For now, human-expert knowledge is superior to AI in many domains, where utilizing human knowledge is extremely helpful (Shoham et al., 2017). Specifically, in regards to food image recognition, allowing users to correct images has been suggested in research developing food image processing algorithms (Hoff et al., 2018). In conclusion, human intervention could be a useful way to interactively enhance AI performance. However, the design of specific interactions and the introduction of human participation has been a neglected research subject. Therefore, we designed interactions to encourage people to participate in the loop and to help an imperfect food-recognizing AI.

3. Hypotheses Development

We conducted two experiments in order to identify the critical factors for encouraging users to participate in the loop. The first experiment was conducted to test the overall impacts of motives of help, while the second experiment was conducted to elaborate on this using the limitations and implications from the first experiment. In this section, we first provide hypotheses for the first experiment, which investigates the overall relations between motives, trust and willingness to help.

3.1. Motives of Help

When an error occurs in an AI system, a reliable correction can be made by its user. Users' help is especially useful in our application domain because our study is about AI for food journaling, which relies on information that users know. In this study, willingness to help is defined as the degree to which users want to help the AI application by adjusting incorrect predictions. Yang, Hsee, and Urminsky (2014) argued that "helping" is a form of prosocial behavior, and measured help by the following question, "How willing are you to do something in order to help another person?" This question was designed to measure a construct defined as "willingness to help". In fact, "willingness to help" is a construct that has been studied in several previous studies (Koster, 2007; Weiner, 1980; Yang et al., 2014). Thus, in this study, "willingness to help" is measured for the evaluation of the level of help provided by users.

This study measures willingness to help using two different types of scales, one subjective and one objective. For the subjective scale, participants were asked to state their willingness to help while using the AI application. In previous studies that used subjective scales, human help was self-reported using questionnaires that asked the participants to rate their degree of willingness to help (Koster, 2007; Weiner, 1980). On the other hand, studies that used objective scales, help was measured by the actual amount of helpful behavior conducted by participants (Berkowitz, 1987; Darley & Batson, 1973; Isen & Levin, 1972). Therefore, the objective scale of willingness to help in this study was based on the number of food pictures that were taken, the amount of monetary reward that was earned as a byproduct of help, and the number of neglecting errors.

We intended to find critical factors to elicit users' help for the AI system by applying a mechanism that encourages help-giving in humans. The hypotheses in this study are based on the theory of human help. Batson et al. (1983) demonstrated that the motivation to help is either egoistic or altruistic. According to the aforementioned study, egoistic motivation is induced when a user wants to relieve their personal distress, while altruistic motivation is triggered by empathy when seeing a person in need. Accordingly, egoistic and altruistic motivations provide the basis for the following triggers which motivate users to help an imperfect AI: utilitarian benefit, empathy, explainable AI, monetary reward, and trust.

3.1.1. Egoistic Motivation

According to Millon et al. (2003), the motive to help can be stimulated when providing help operates as a means to achieve self-benefit. An egoistic motivation for giving help is elicited when a user believes that they can benefit from their helping behavior. For instance, donation is influenced by the perception of self-benefit in both direct and indirect forms (Amos, 1982). In other words, donation tends to occur when a donor believes that they will benefit as a result of their deed.

Similarly, when people feel that they have benefitted from someone, they cooperate in return (Fehr & Gächter, 2000). Benefit is, in fact, a basic motive that moves people and leads human beings to perform actions in order to avoid punishment or attain reward (Vallerand, 1997). For instance, perceived benefit increases the likelihood that a recommendation will be accepted (Becker & Maiman, 1975). In order to trigger egoistic motivation, this paper provided utilitarian benefit in the first study and explainable AI and monetary reward in the second study. It is posited that providing a benefit to correction would motivate users to participate in the loop.

The utilitarian benefit of using the application was designed to make participants perceive self-benefit. A number of marketing researchers have demonstrated that utilitarian benefit is the main benefit that influences consumer behavior (Babin, Darden, & Griffin, 1994; Chandon, Wansink, & Laurent, 2000; O'Brien, 2010; Pallas, Mittal, & Groening, 2014). In this study, we divided utilitarian benefit into the benefit of explanation itself and the benefit of improved system performance, which was conveyed via an explanation. The former was operationally defined as the benefit of explainable AI. According to Gunning (2017), explainable AI "provides end users with an explanation of individual decisions, enables users to understand the system's overall strengths and weakness, conveys an understanding of how the system will behave in future, and perhaps demonstrates how to correct the system's mistakes." For instance, a current AI system based on a deep learning algorithm does not explain how 'cat' is predicted from an image containing a cat. In comparison, explainable AI justifies its prediction by showing users that a certain picture is a cat because features such as fur, whiskers, and claws were detected. Explanations provided by explainable AI makes the AI system more transparent and are expected to reduce the uncertainty around AI, which can be considered a utilitarian benefit to users. In fact, explaining the results of an intelligent decision aid has itself been shown to be perceived as a benefit by users (Arnold, Clark, Collier, Leech, & Sutton, 2004). The latter form of benefit shows a function that allows users to accumulate points as they take pictures and correct the predictions of AI; these points can be used to buy a larger database that improves prediction quality. It is informed to the participants that when prediction quality increases, prediction time decreases, and they can accomplish their tasks faster. Providing information about the benefits of their behavior motivates people to act (Delgado, 2007; Duflo & Saez, 2003). To sum up, explainable AI reduces the uncertainty and provides utilitarian benefit, and the explanation of the AI's improved performance allows users to enhance prediction quality, which can be perceived as a benefit for the users. The utilitarian benefit from explainable AI and the utilitarian benefit of the AI's improved performance are the two aspects of utilitarian benefit used in this study to trigger the egoistic motivation to help. Since prior studies have shown that egoistic motivation from perceived benefit increases willingness to help, we used the utilitarian benefit as the benefit that increases users' willingness to help the AI. Therefore, an application with utilitarian benefit was developed to test the following hypothesis:

H1a. *Utilitarian benefit will have a positive effect on users' willingness to help AI.*

3.1.2. Altruistic Motivation

Many studies have emphasized the importance of emotion in help, especially empathy, which has been shown to increase the probability that help will be given (Eisenberg & Miller, 1987). The attribution-affect-action model constructed by Weiner (1980) theorized that people first recognize the context, then feel emotion accordingly, and take action as a consequence. For helping behavior, during the attribution phase, people analyze the context of others in need based on locus, stability, and controllability. When emotion followed the attribution stage, helping behavior was significantly affected, compared to conditions when emotion was not evoked. In summary, the decision to help entails a certain process, and stimulation of emotion is a significant factor. Batson et al. (1981) implemented an experiment that revealed empathy lead to the altruistic motivation of help. When participants felt high levels of empathy, even when they could easily avoid the situation, they helped as much as when it was difficult. On the other hand, when low levels of empathy were induced, participants would help only when it was hard to evade. Therefore, empathy is shown to be positively associated with willingness to help for other human beings.

A technological cue, such as VR, is capable of being an effective means to trigger empathy (D. Shin, 2018). When a technological cue stimulates empathy, its users experience more realistic and empathic technology, leading to an increased continuance intention (D. Shin & Biocca, 2018). Since help for AI can be implemented when users continue their usage, continuance intention is an essential factor of willingness to help. In summary, as this paper aims to study the components that positively affect trust and willingness to help for AI users, an AI system that incorporates empathy is designed. Humans feel empathy towards an inanimate object when it is similar to a human in some way (Mori, 1970). This implies that an AI designed to induce empathy should contain a sufficient number of human traits. Computer systems intending to imitate humans should think and even feel like a human, implying that they should show emotions (Martınez-Miranda & Aldea, 2005). Virtual agents have successfully elicited empathetic behaviors from users by interacting with them (Paiva et al., 2005). In conclusion, inanimate objects with humanlike features elicit empathy, which generates the continued use of technology and the altruistic motivation of help. The elevated continuance intention and altruistic motivation of help are expected to raise the willingness to help an AI. Because inanimate objects with humanlike features elicit empathy from users, this study utilized an AI agent character with various facial expressions to garner empathy from the participants, leading to an increased level of willingness to help the AI. Consequently, a humanlike AI agent was utilized in the experiment to test the following hypothesis:

H1b. *Empathy will have a positive effect on users' willingness to help AI.*

3.2. Trust

Trust is a critical concept in system design because an imperfect AI is likely to be rejected unless a reasonable level of trust is generated between humans and the system (Muir, 1987). When the accuracy of a system is not guaranteed, users' experience of AI is ruined (Dill & Rabin, 2013). At present, users cannot trust AI applications completely, owing to their performance flaws. Trust is a crucial part of the user experience as it affects users' thoughts and feelings towards the system (J. D. Lee & See, 2004). For instance, in the field of social commerce platforms, trust was shown to have a positive effect on customer behavior (D.-H. Shin, 2013). For social media, trust was a critical component with regard to privacy issues in that users were less reluctant to reveal their private information on social media when they had enough trust in the platform (D.-H. Shin, 2010). People tend to rely on automation that they trust (J. Lee & Moray, 1992). When trust is not generated, people do not accept decisions made by the system (Muir, 1987). In other words, it is important to design trustworthy interactions between humans and AI to provide a positive user experience. This study argues that trustworthy AI can be achieved through utilitarian benefit and empathy.

3.2.1. Trust and Utilitarian Benefit

In this study, the utilitarian benefit construct is composed of the explanation itself and improvement in the system's performance, which is conveyed through explanation. In other words, explanation is the source and means of the benefit. These aspects of the utilitarian benefit construct that are related to explanation will have more effect on the trust construct. Miller (2017) noted that social and psychological principles of human explanation could be applied to an intelligent system. Sense-making is an innate ability of humankind that allows us to comprehend our experiences (Klein, Moon, & Hoffman, 2006). Sense-making involves user's supposition of how the AI makes its predictions. The accuracy of sense-making depends on the individual's preliminary knowledge of AI, and it can be utterly different from the actual AI process. In other words, there may be a considerable gap between the actual algorithm and a human's sense-making of an AI algorithm. When users do not have enough information about a system, they assess its reliability subjectively (Muir, 1987). Incorrect or not, users of a system form their trust based on their own sense-making process. By providing an explanation of an AI process, the gap between users' own sense-making and the actual AI algorithm is reduced and users' understanding increases accordingly, which leads them to trust the AI.

Understandability is related to human trust in the system. The extent to which a user understands the AI is a decisive factor in the formation of trust (Ribeiro, Singh, & Guestrin, 2016). Prior studies have demonstrated that increasing transparency reinforces trust. Kulesza et al. (2015) showed that higher transparency, compared to the traditional black box, led to greater understanding and allowed users to create more accurate corrections for the system. Mercado et al. (2016) demonstrated that a participant's performance, trust, and usability were higher for systems with higher transparency. When the system made an error, users' distrust increased rapidly, but an explanation of the reason why the error occurred enabled the rise of reliance (Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003). In addition, explaining why the system operated in a certain way increased understandability for users, leading to a higher level of trust (Lim & Dey, 2011; Lim, Dey, & Avrahami, 2009). Based on prior studies that demonstrated the importance of understandability in human trust, this study developed the utilitarian benefit construct, which consisted of explainable AI and performance improvement to increase users' understanding of the AI system. The increased understanding by the users was expected to close the gap between the users' own sense-making processes and the actual AI algorithm, thus increasing the users' trust in the system. In summary, when explanations of the system's internal process were included, they increased system transparency and enhanced the users' understanding of the AI system. When users understand why the system has made incorrect predictions, trust in the system can be maintained. Therefore, we posited the following hypothesis:

H2a. *Utilitarian benefit will have a positive effect on trust between a user and AI.*

3.2.2. Trust and Empathy

Empathy forms trust between people, systems, and institutes. Aside from competence and reliability, empathy has received attention as a means to form trust in the field of customer relationships. In precedent studies, empathy was shown to form trust between a customer and a service provider (Coulter & Coulter, 2002), a patient and a physician (Hojat et al., 2010; Kim, Kaplowitz, & Johnston, 2004), and in a transaction of automobiles (Spaulding & Plank, 2007). In the field of human-computer interaction, a relationship developed online relied on empathy to form interpersonal trust in that when a supportive response such as empathy was provided, interpersonal trust was achieved (Feng, Lazar, & Preece, 2004). Following this principle, the AI agent in this study expressed its emotion in accordance with the rate of users' help and prediction accuracy. If the rate of users' help and prediction accuracy were low, the AI agent expressed negative emotions and vice versa. The relationship between empathy and trust between people are applicable to human-machine trust (J. Lee

& Moray, 1992). Rendering empathy in the users of a system with virtual agents was shown to be viable (Paiva et al., 2005). In short, virtual agents evoke empathy, and empathy is a fundamental source of trust between people. The association of empathy with trust between human beings is applicable to the relationships between humans and AI, because humans tend to treat computers as other human beings (Nass & Moon, 2000). Based on these findings, an AI agent with emotional factors was designed to induce empathy, leading to enhanced trust in the AI. Thus, the following hypothesis was proposed:

H2b. *Empathy will have a positive effect on trust between a user and AI.*

3.2.3. Trust and Willingness to Help

This study explored the components that positively affect AI users' willingness to help. Trust is an essential precursor of prosocial behavior because prosocial behavior, including help, requires personal sacrifice. When users believe that their sacrifice will be compensated, which is largely affected by the depth of trust, their cooperation is induced (Twenge et al., 2007). Trust has been shown to affect prosocial and cooperative behaviors in a number of studies. The field of brain science has revealed that oxytocin, the hormone that increases trust between humans, also affects a person's willingness to bear social risks in order to conduct prosocial behavior (Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005). In the field of marketing, studies have shown that trust has an impact on the relationships between customers and companies, and trust increases the probability of initiating a relationship (Gounaris, 2005). McAllister (1995), the author of a study which was used as a basis for the development of the human-computer trust (HCT) scale, argued that trust positively affects assistant citizenship behavior, referring to help-giving behavior in a workplace. This study adopted factors that were shown to be effective in triggering trust between human beings in a variety of fields, because trust was shown to be positively associated with helping behavior. Since the influence of trust on human relationships can be applied to human–machine interaction (J. Lee & Moray, 1992), we utilized four factors that were shown to increase humans' willingness to help: utilitarian benefit, empathy, explainable AI, and monetary reward. These constructs were designed to trigger trust, which was hypothesized to mediate the effects of these factors on the users' willingness to help. Based on this, we developed the following hypothesis:

H3. *Trust between a user and AI will have a positive effect on users' willingness to help AI.*

The research model of Study 1 established based on literature review, is illustrated in **Figure 1.**

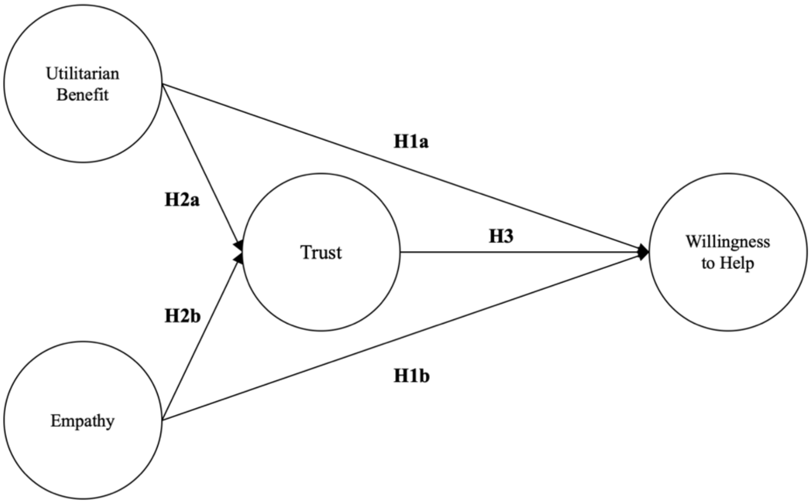


Figure 1. Research model of Study 1

4. Study 1

4.1. Method

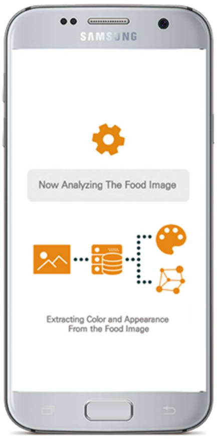
An experiment was conducted to investigate the effects of utilitarian benefit and empathy on the users' trust and willingness to help. In order to test our hypotheses, a 2 (high vs. low utilitarian benefit) × 2 (high vs. low empathy) between-subjects experiment was designed. The materials for each cell were developed using an Android smartphone application.

4.1.1. Materials

Materials for Utilitarian Benefits

We conducted a qualitative pre-test to understand the way people think about how AI analyzes food images. A prototype application was developed for participants to experience food image recognition using an AI mobile application. One-hour interviews were conducted with eight participants. After each participant used the prototype food recognizing AI application, we asked them what processes they thought were used to generate the prediction from the image. The prototype application did not explain how the result was predicted, so participants had to guess its method. All of the participants said that they believed that the image's color and shape were used. According to this pre-test result, we designed an experiment stimulus for the explainable AI by describing the process of recognition including color and shape as the prediction criteria.

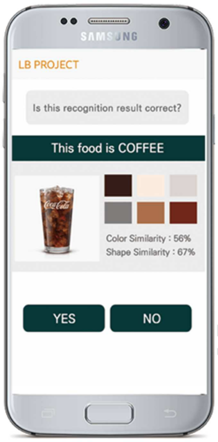
The first part of the explanation covered the food recognition algorithm’s processes and the features that it extracts from the captured food images (**Figure 2**). In addition, the application indicated the similarity between a captured image and an image in the AI database in terms of a percentage of color and shape, as shown in **Figure 3**. Meanwhile, we designed the utilitarian benefit condition to trigger an egoistic motivation to help. In order to manipulate participants into having an egoistic motivation to help, it was necessary for participants to think that they benefitted from the application when they helped it. Each time participants corrected the AI prediction; 50 points were accumulated. By accumulating points, participants could buy datasets that increased the number of food images they had access to, which was explained as a way to improve the performance of the application.



Now Analyzing The Food Image

Extracting Color and Appearance From the Food Image

Figure 2. Screenshot of the application. It explains that the image recognition algorithm is operating, and the analysis standards, which are color and shape.



Is this recognition result correct?

This food is COFFEE

Color Similarity : 56%
Shape Similarity : 67%

Figure 3. The percentage of color and shape similarity is shown in the application.

The flow of the application with the utilitarian benefit is shown in **Figure 4**. In the main page (i) the icon at the top of the interface rotated to indicate that the image recognition algorithm was operating. The performance of the application was shown on the main page. The default performance was ‘medium’. The more the participants helped the system, this increased from ‘medium’ to ‘high’ performance. After taking a picture, it took a few seconds for the recognition process to be completed. During that time, the system explained that the algorithm was working, extracting the color and shape from the image, as shown in (iv). In (v), the recognition result was presented with the color and shape similarity. If the participants chose not to help by tapping the ‘NO’ or ‘SKIP’ button, the trial was completed without giving 50 points to the users. In contrast, if they helped to correct the result, 50 points were accumulated. The saved points and current database set were shown on the main page. If participants tapped the “BUY DATABASE” button on the left side of the interface, they could see more extensive database sets that could be bought using the points. In contrast, participants without utilitarian benefit did not receive any details about the progress or performance. Once they took a picture, they were led to the results page that displayed only the prediction. Moreover, database related functions were not provided, meaning that the main page was composed solely of a camera button.

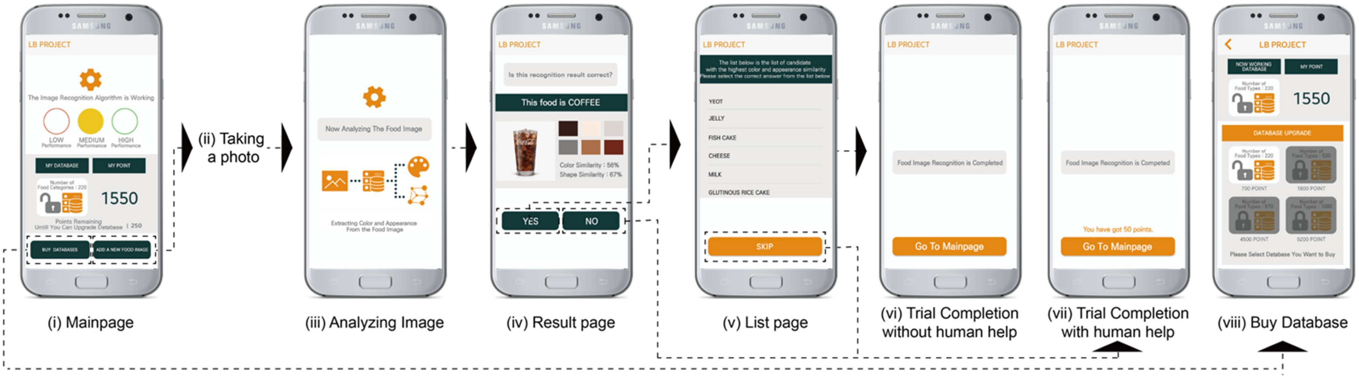


Figure 4. Workflow of the application for utilitarian benefit based on egoistic motivation.

Materials for Empathy

We induced participants to have higher empathy by showing them an AI-agent character named ‘Eggy’, which used a colloquial form of language to present what was going on in the application. Eggy revealed its emotion using facial expressions and words such as tiredness, pleasure, and embarrassment (see **Figure 5**). People typically feel empathy towards other human beings, hence, when an inanimate object manifests an acceptable level of humanness, affinity arises (Mori, 1970). Misselhorn (2009) argued that people can feel empathy toward inanimate objects through imaginative perception. In fact, people, even children and infants, instinctively react to other people’s emotional expressions (Meltzoff & Moore, 1983). Based on the fact that facial expression is a primary means of emotional expression in human beings that effectively leads to empathy, we designed an avatar with a variety of facial expressions in order to trigger empathy.

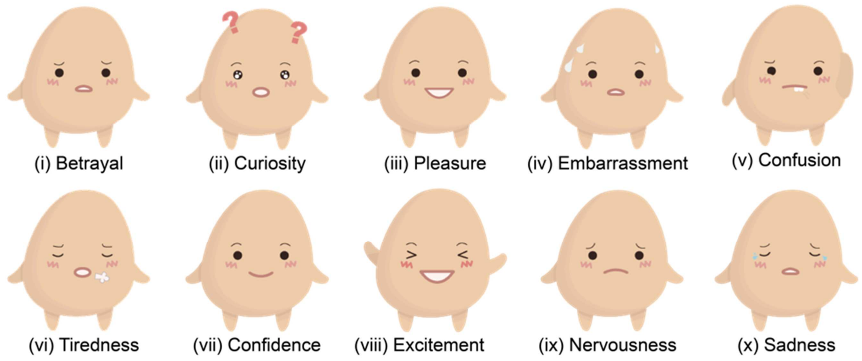


Figure 5. The AI-agent character ‘Eggy’. Eggy expressed ten types of emotion which are betrayal, curiosity, pleasure, embarrassment, confusion, tiredness, confidence, excitement, nervousness, and sadness.

Image recognition algorithms calculate scores based on the degree to which the AI believes its recognition is close to the correct answer (Bezdek, Keller, Krisnapuram, & Pal, 1999). In line with this characteristic of AI image analysis, we designed the positive emotion of the character to be associated with the confidence level of prediction. Furthermore, immediately before the participants started to use our application for the experiment, we showed them a vivid story of how the character was born and why she came to work for the food recognition application.

The procedure with empathy was designed as shown in **Figure 6**. On the main page, participants could check Eggy's current mood, which was bad, neutral, or good. At first, Eggy's state was "neutral"; as participants gave more help to the application, Eggy's mood improved. Participants took a picture of prepared food images and waited until recognition finished while (iii) was displayed. The recognition result was then shown, as in (iv). Participants decided to correct Eggy's prediction at will. If they chose not to correct it, or to skip the correction after entering (v), Eggy expressed her disappointment as in (vii). If participants decided to help the application, Eggy expressed her happiness as in (vi) after the error-handling was completed. In contrast, participants without empathy materials were not provided with any aspects related to Eggy. Thus, the main page contained one button for adding new food images and after the participant took a picture, the result page was shown directly.

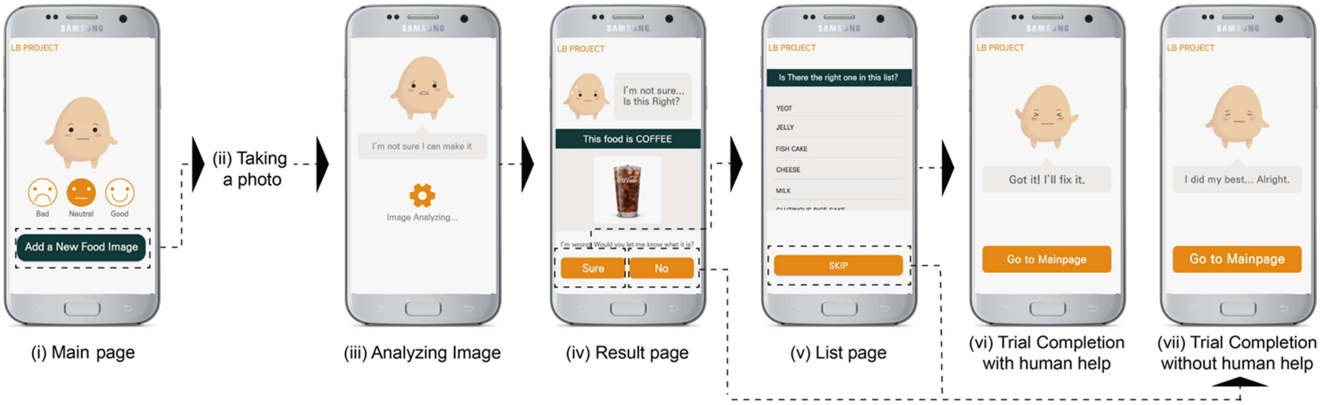


Figure 6. Workflow of the application for arousing empathy with altruistic motivation

After the adjusting process is completed, the adjusted data can be utilized in the HITL system. Here, HITL refers to a system that allows users' corrections for incorrect food image predictions, which are instantly reflected in the AI algorithm. After the human correction is taken into the system via user interface that allows human input (see Fig. 4(v), Fig. 6(v) and Fig. 9(iii)), the HITL algorithm trains on human-labeled data to enhance the algorithm. Specifically, when the accurate human inputs replace incorrect predictions, the accuracy of AI is increased by the elimination of classifier errors (Fails & Olsen Jr, 2003).

4.1.2. *Procedure*

Four prototype applications for the two by two experiment groups were developed with Android OS and they were installed on a test smartphone, Galaxy Note 4. This study was implemented with Wizard of Oz experiments, and the participants were given a smartphone with an activated prototype application. A total of 20 printed food images were presented to the participants. All of the participants experienced the same recognition performance. If the actual accuracy for each participant was different, it would have been impossible to compare each participant's trust and help because the difference in actual performance would have biased them. Therefore, the experimental application was designed to show the same result according to the preset scenario, no matter what kind of pictures were captured. With the equivalent recognition performance, we studied the effect of two research stimuli, utilitarian benefit and empathy. The dependent variable, willingness to help, and the mediating variable, trust, were subjectively assessed using several questionnaires. In terms of inducing users' help, the recognition performance improved as the trial number increased. Additionally, the latency of recognition gradually decreased as participants continued to use it. Participants conducted 20 predictions, and were notified that if the corrections to the AI predictions were completed unreliably, then the participation reward would only be partially paid. Consequently, all of the participants adjusted incorrect predictions during the 20 predictions in the trial. Once 20 rounds of recognition were completed, questions were provided regarding the manipulation checks.

4.1.3. *Measurements*

All of the questionnaires were measured on a 5-point Likert scale ranging from 1 (strongly disagree/low) to 5 (strongly agree/high). The following questionnaires relating to utilitarian benefit and empathy were utilized for manipulation and hypotheses testing. Specific questionnaire items can be found in **Appendix A.**

Utilitarian benefit

Given that utilitarian benefit construct included an explanation of the performance improvement of the AI and its benefits, the questionnaires focused on the performance advancement and its perceived benefits. Participants were asked to rate the perceived performance advancement and whether it met their expectations.

Empathy

Batson's (2002) empathy adjectives were used to assess the level of the participants' stimulated empathy. Four adjectives were provided and the participants had to indicate the extent to which they felt that way when using the application. Each word was translated to convey the meaning as closely as possible.

Trust

The human-computer trust (HCT) scale was utilized to measure the participants' trust. Questionnaires for cognition-based trust between humans and computers were used (Madsen & Gregor, 2000). Human-computer trust (HCT) is based on McAllister's (1995) study of human-human trust, which is defined as the degree to which users trust and decide to use AI. (Madsen & Gregor, 2000). Cognitive trust in HCT, which is based upon users' rationale, was used to measure trust in this study. Perceived understandability, technical competence, and reliability affect cognition-based trust. Understandability means that human users are able to make a mental model and predict the working processes of the AI, technical competence reflects the accuracy of tasks done by AI, and reliability represents the consistency of the performance of the AI. Given that this study covers trust in an AI system, a trust construct developed to measure human trust in an AI system is appropriate.

Willingness to help

Two questions were presented to measure the participants' willingness to help: "When the application asked me to correct the food recognition result, i) I was willing to help it. ii) I was annoyed. (reverse-coded)". Willingness to help questionnaires were created based on prior works that measured participants' willingness to provide help (Koster, 2007; Weiner, 1980; Yang et al., 2014).

4.1.4. Validity and Reliability

A total of 14 questions were developed. Of these, 12 questions were from validated former studies (Batson et al., 2002; Koster, 2007; Madsen & Gregor, 2000; Weiner, 1980; Yang et al., 2014) and the rest were created to measure the explanation construct. All of the questionnaires were translated into Korean by the authors, then they were co-reviewed by three other researchers in HCI, and finally they were revised by an HCI expert. The test results of reliability, convergent validity and discriminant validity can be found in **Appendices A** and **B**, respectively. Verification tests were conducted with Smart PLS 3.0. The factor loadings of each item were above 0.70, except for one item of explanation (0.648). The item which did not meet the validity standard will be explained as a limitation of Study 1 in the discussion. Cronbach's alphas and composite reliabilities for all the factors exceeded 0.70, indicating that the reliability of the research is acceptable. In addition, the discriminant validity of Study 1 was found. Square roots of the average variance were extracted (diagonal elements in **Appendix B**) and were shown to be larger than the correlation between one latent variable compared to the other variables (off-diagonal elements).

4.2. Results

4.2.1. Participants

Participants were recruited through several online communities frequently visited by younger generations (20 to 30) who are familiar with smartphone applications. A total of 76 participants (29 males; 47 females; average age of 24) joined in our experiment. Participation rewards worth 8 USD were paid once the experiment was completed. All of the participants were educated to undergraduate or graduate school level. Out of the 76 participants recruited, 3 participants' data were eliminated. We drew a box-plot with SPSS 23 and the values that did not fall into the interquartile range multiplied by 1.5 were considered to be outliers. Consequently, experimental data from 73 participants (28 males; 45 females) was analyzed in this study. In addition, 16 participants in the control group, 19 in empathy condition, 17 in utilitarian benefit condition, and 21 in both empathy and utilitarian benefit conditions were analyzed. No statistically significant difference was found between different gender groups.

4.2.2. Manipulation Check

Two manipulation tests were implemented. The level of a participant's attention was measured using a task that asked them to separate phrases displayed in the application from those which were not. Two independent variables, utilitarian benefit and empathy, were shown to be well manipulated. The utilitarian benefit manipulation was measured through the benefit and explanation related questions, and the two-tailed t-test showed that there was a significant difference (t(71) = 2.183, p = 0.033*) between groups with (M = 3.50, SD = 0.65) and without the utilitarian benefit (M = 3.11, SD = 0.85). A t-test for the second factor, empathy, was also conducted. There was a significant variation in the empathy adjective scale points (t(71) = 4.231, p = 0.000***) between the group with (M = 2.73, SD = 1.07) and the group without empathy (M = 1.80, SD = 0.79). An overview of the manipulation test is illustrated in **Table 2.**

Table 2. Manipulation check results for study 1

Independent variables	**Mean (std. dev.)**		**t-statistic**		
	High	**Low**	**Mean difference**	**t-statistic**	**Significance level**
Utilitarian benefit	3.50 (0.65)	3.11 (0.85)	0.389	t(71) = 2.183	p = 0.033*
Empathy	2.73 (1.07)	1.80 (0.79)	0.924	t(71) = 4.231	p = 0.000***

*p < 0.05, **p < 0.01, ***p<0.001.

4.2.3. Hypotheses Testing (Structural Equation Model)

Structural equation modeling (SEM) was used to test our hypotheses (Anderson & Gerbing, 1988; Bagozzi & Yi, 2012). We chose this because SEM is appropriate for psychometric variables and more complex models than traditional regression analyses, including mediating variables, such as trust in our study. The model testing was conducted using the partial least squares structural equation modeling (PLS-SEM) tool in SmartPLS3 (Hair Jr, Hult, Ringle, & Sarstedt, 2016; Ringle, Wende, & Becker, 2015). To calculate the t-statistics, we used bootstrapping with 200 resamples (Chin, 1998). **Figure 7** shows the path coefficients for the whole model and the associated statistical significance. The PLS results showed that utilitarian benefit had an indirect effect on willingness to help, while empathy was shown to have a direct effect. Of H2a and H2b, only H2a was significantly supported with a path coefficient of 0.629**. H3 was supported (path coefficient = 0.286**), indicating that utilitarian benefit had an indirect effect on willingness to help. The direct effect was only shown for empathy. H1b had a significant coefficient of 0.346**, which suggests that empathy has a positive effect on users' willingness to help. In contrast, H1a was not supported (path coefficient = 0.061). The structure of the hypotheses and PLS results are shown in **Figure 7** and **Table 3**.

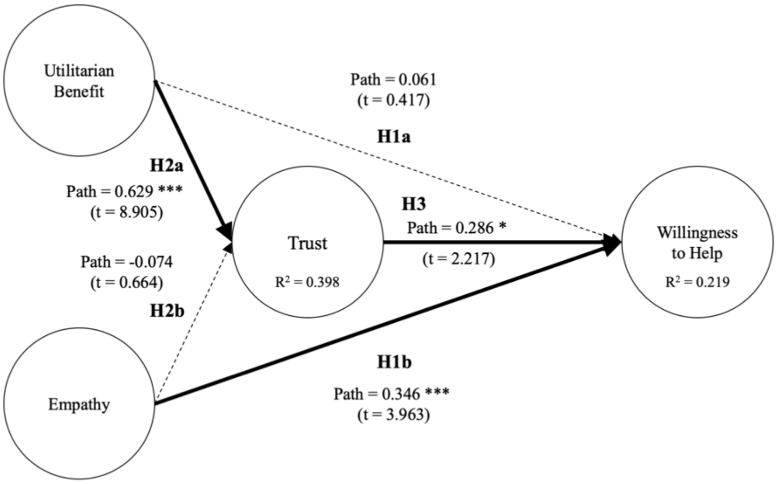


Figure 7. Structural model of Study1

Table 3. Partial least square results of Study 1

Path (Operational Variables)	**Hypotheses**	**Path Coefficient**	**t-statistics**	**p-value**	**Hypothesis Supported**

Utilitarian Benefit → Willingness to help	H1a	0.061	0.417	0.677	-
Empathy → Willingness to help	H1b	0.346	3.963	0.000***	supported
Utilitarian Benefit → Trust (HCT)	H2a	0.629	8.905	0.000***	supported
Empathy → Trust (HCT)	H2b	-0.074	0.664	0.507	-
Trust (HCT) → Willingness to help	H3	0.286	2.217	0.028*	supported

*$p < 0.05$, **$p < 0.01$, ***$p<0.001$.

4.3. Discussion

It was posited that utilitarian benefit and empathy would be positively associated with willingness to help. The PLS results revealed that empathy had a direct effect on willingness to help, whereas utilitarian benefit had an indirect effect on willingness to help, mediated by trust. However, there were several limitations to this study, which led to the second study of this research.

First, the hypothesis analysis with PLS demonstrated that the effect of empathy on trust was insignificant (H2b) and that the utilitarian benefit did not have a direct effect on willingness to help (H1a). Regarding the former issue, it was deemed that the robust direct effect of empathy on willingness to help was responsible for the insignificant result of H2b. In other words, empathy is effective in inducing willingness to help regardless of trust. On the other hand, there are two underlying causes for the latter issue: utilitarian benefit not being considered as a direct advantage and the compounded design of the utilitarian benefit construct. In the in-depth interview, participants tended to respond positively to the explanation of providing utilitarian benefit. One of our participants said: *"As a user, helping this* application *also helps me. So if I'm asked for help, I think I would give help."* [P65] In addition, it was also mentioned that: *"Eventually, if I use this* application *for a long time, enhancing performance would be helpful for me, so I will correct the results as long as I'm still using it."* [P29] The responding participants factored in the long-term advantages that they would gain when applying it in their real lives.

Therefore, it is possible that the participants did not regard the utilitarian benefit provided by our experiment as being directly advantageous during the experiment, resulting in an insignificant effect of H1a. Moreover, participants experienced explainable AI and the benefit of performance enhancement at the same time. The utilitarian benefit construct contained two different sub-constructs, explainable AI and the explanation of the utilitarian benefit of performance advancement. This problem was illustrated by the factor loading of one item that was used to measure utilitarian benefit, which was lower than 0.70 (0.648). We believe that this was because one measure was about the perceived utilitarian benefit and the other two were about the explanation of performance enhancement. Even though explainable AI and the explanation of the utilitarian benefit were both intended to be a benefit in the form of an explanation, explainable AI is closer to the concept of an explanation, while the explanation of the benefit is closer to the concept of a benefit. Thus, it is difficult to tell which factor was responsible for the indirect effect. In other words, it is possible that there was no direct effect of utilitarian benefit because it was a compound of explainable AI and performance enhancement. So, in Study 2, we separated the explanation from the utilitarian benefit and changed utilitarian benefit to a monetary reward, which is a more powerful type of benefit.

Another problem with study 1 was that, in contrast to the PLS results, the ANOVA results did not reveal any difference in willingness to help between the conditions with and without empathy. It can be said that this result occurred because people who are prone to feeling empathy, regardless of the empathy condition, reported greater willingness to help. Another explanation is that a number of participants did not provide straightforward answers about their willingness to help. In the post-interview, when the participants were asked why they had helped the AI to correct wrong predictions, 13 participants said that it was because they wanted to enhance the performance of the application, not for their own benefit but for the researchers or the developers of the application. Moreover, this study notified participants before the experiment that they would be taking part in a usability test for an application prior to launch. Consequently, the participants seemed to be conscious of the experimenter as they believed that the application was a business item of the experimenters. Thus, it may have been difficult for the participants to give a negative response. In fact, there was a tendency for the participants to give high points for willingness to help, resulting in no significant differences between the groups. All of the groups except the control group gave average points over 4. To solve this problem, in Study 2 we substituted the subjective scale with an objective scale, in order to measure the users' willingness to help. The objective scale measured participants' behavioral features and was expected to reveal the participants' actual willingness to help.

The third limitation is that, during the first experiment, we tried to measure the participants' willingness to help with an objective scale. Unfortunately, the results were not adequate for statistical analysis, because there were not enough tasks to cause the participants to feel weariness. For the first study, we prepared 40 tasks and allowed the participants to fix the results by choosing the correct answers from a list. However, we did not realize that these tasks were relatively easy and able to be completed with minimal effort. Only 8 of the 73 participants took less than 40 prepared pictures, and 3 participants decided not to adjust the incorrect predictions, even those that resulted from a mistake. This problem was addressed in Study 2 by both increasing the number of tasks and the cognitive load of each task, of which the latter was achieved by requiring users to type the entirety of each answer for the correction.

The compounded design of the empathy construct is the last limitation. The empathy construct was facilitated by an anthropomorphized AI agent. The compounded effect of the user interface design of an AI agent and empathy made it difficult to determine the individual impact of empathy. In other words, it is possible that the users were willing to provide help not just because they felt empathy from the AI agent but because the AI agent's user interface had a considerable impact on the users. In order to untangle this issue, we decided to design the prototype applications with empathy for all of the conditions in Study 2. There are two reasons for not dividing the empathy construct into an anthropomorphized agent and empathy but rather applying the AI agent to all conditions: the results of Study 1 showed that empathy is a robust source of willingness to help, and prior studies and existing services incorporated anthropomorphized characters to trigger empathy or to provide a positive user experience.

To be specific, for Study 1, the PLS results showed a strong association between empathy and willingness to help, and many of the participants stated that the AI agent was their reason for helping. One participant [P1] expressed his emotion as follows: *"It was so cute... and whenever the character made an incorrect prediction, crying and asking if she's right, made me naturally correct the result … This made me feel like she's my friend, so I could do it easily without getting annoyed."* Another supporting response was: *"I fixed the results because of Eggy, without her, I wouldn't have wanted to do it much more."* [P74].

Moreover, prior studies applied an embodied agent, elicited empathy and yielded positive outcomes such as increased willingness to continue learning (Chen et al., 2012) and overcoming social communication disorders (Johnson et al., 2018). These positive results may have been produced by the empathy aroused in the users or by the user interface design itself. The existing services and applications contain characters as the system agents (Pearl, 2016), regardless of the exact cause of the enhanced result because, either way, users can have a positive experience. Therefore, to reflect the trend of including a system agent character, we have set empathy condition as the control variable and applied embodied character in all experimental groups in Study 2.

5. Study 2

Based on the limitations and insights from Study 1, we developed the hypotheses, measurements, experimental processes, and materials for Study 2.

5.1. Background and Hypotheses Development

The research model in Study 2 separated the utilitarian benefit condition from Study 1 into two variables, changing four of the hypotheses (H1a, H1b, H2a, and H2b). Other adjustments made in Study 2 were the establishment of the empathy construct as a control variable and inclusion of Eggy in all of the experimental conditions. Because empathy is a control variable, the individual effect of empathy was not measured. All of the research hypotheses in Study 2 are shown in **Figure 8**. Since our explanations in section 3 already provided detailed backups for most hypotheses of Study 2, only additional backups to support the changed hypotheses are provided below.

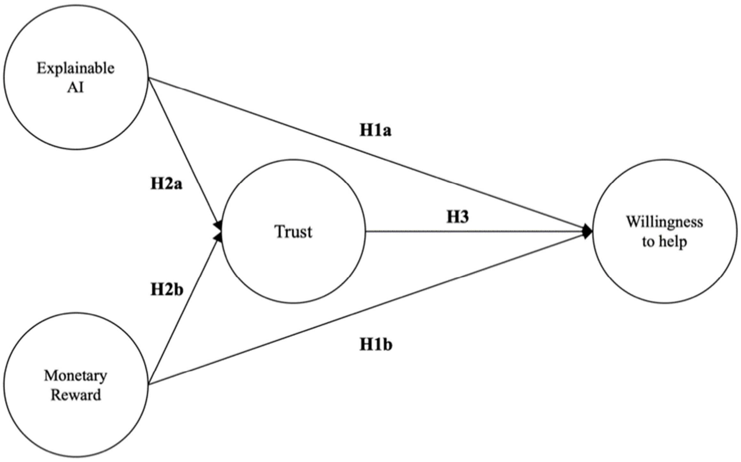


Figure 8. Research model of Study 2

5.1.1. *The Effect of Explainable AI on Willingness to Help and Trust*

Given that the explanation condition no longer included a description of benefit, the explanation was operationally defined as explainable AI. Explainable AI is posited to have a direct effect on willingness to help, not because of self-benefit, but because it explains the reason for incorrect predictions. The attribution-affect-action model suggests that people evaluate the context of others in need and then make a decision whether or not to provide help (Weiner, 1980). When people decide to help, it is based on two factors. The first factor is locus, or whether the cause of the problem is internal or external to the person in need. The second factor is controllability, or whether the cause is controllable by the person in need. Explainable AI not only extracts the prediction but also provides the basis on which it was made. The explanations provided in this study demonstrated how the predictions were calculated and enabled users to understand that incorrect predictions by the AI were due to external influences and limited controllability. Specifically, the explainable AI in this study presents its confidence level with regard to color and shape similarity, as well as colors extracted from an image. The extracted colors are not all correctly predicted due to confusion with colors present in the background of an image. For example, in images of food, the desk behind the food may be extracted alongside the food. This information is presented to the users when the explainable AI shows them the extracted colors. If the extracted colors are not provided, the users cannot understand the reason for the incorrect predictions, leading to their own sense-making of the result. Moreover, the confidence level for the prediction similarity shows the limited controllability of the AI, which helps users understand the AI algorithm's capabilities and limitations. Users can expect as much accuracy as the confidence percentage level allows. Combined with the attribution–affect–action model of the decision to help, the explainable AI provides reasons for the incorrect predictions and is capable of increasing willingness to help. In terms of locus, these explanations present external factors that led to the incorrect prediction, such as an imperfect input image. With regard to controllability, the AI algorithm's limited control is presented by means of the similarity confidence level. Thus, these features of an explainable AI were incorporated into Study 2 and we hypothesized that:

H1a. *Explainable AI will have a positive effect on users' willingness to help AI.*

In addition, as mentioned in section 3 of this paper, explainable AI increases the transparency and understandability of the system, and affects the users' trust, supporting hypothesis H2a.

H2a. *Explainable AI will have a positive effect on trust between a user and AI.*

5.1.2. The Effect of Monetary Reward on Willingness to Help and Trust

Benefit is a basic trigger that motivates people (Vallerand, 1997), including helping behavior (Millon et al., 2003). Accordingly, the decision to help is largely affected by the self-benefit that can be achieved from one's altruistic behavior (Fehr & Gächter, 2000). Monetary reward is considered a powerful form of benefit that motivates people (Libera & Chelazzi, 2006). As illustrated in section 3, because benefit increases users' willingness to help, monetary reward is hypothesized to be positively associated with willingness to help. Thus, a monetary reward was provided as a benefit for participants in Study 2, and the following hypothesis was established:

H1b. *Monetary reward will have a positive effect on users' willingness to help AI.*

According to Williamson (1993), trust based on calculation is generated when the other party is expected to behave beneficially enough for them to cooperate. Earlier studies have shown that a feedback mechanism in an e-commerce platform that assesses the credibility of agents fosters trust in the platform and other agents (Ba & Pavlou, 2002; Ba, Whinston, & Zhang, 2003). This is because positive ratings from the feedback mechanism, which can result in increased transactions, are considered to be a benefit, while negative ratings are considered a threat. A robust form of benefit that motivates people is monetary reward. Economic gain is a fundamental human motive, even for altruism (Simon, 1993). Chung & Kim (2009) demonstrated that, although people usually tend to avoid hazardous facilities, when an economic benefit is expected, trust and acceptance of the facilities increase. In conclusion, monetary reward is a robust form of benefit that motivates people to act and feel in certain ways, including trust. Therefore, as part of this study, a monetary reward construct was developed to test the following hypothesis:

H2b. *Monetary reward will have a positive effect on trust between a user and AI.*

In addition, the hypothesis related to the association between trust and willingness to help in Study 1 was utilized again in Study 2 to investigate the mediating effect of trust.

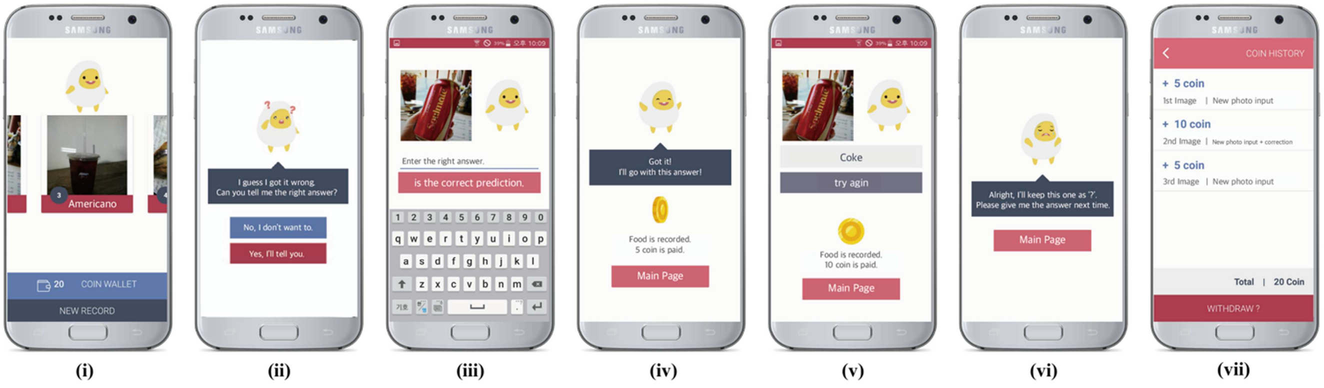
H3. *Trust between a user and AI will have a positive effect on users' willingness to help AI.*

5.2. Method

An experiment was conducted to examine the effects of explainable AI and monetary reward on users' trust and willingness to help. Two factors were manipulated in a 2 (high vs. low explainable AI) × 2 (high vs. low monetary reward) between-subjects design. The materials for each cell were developed using an Android smartphone application.

5.2.1. Experiment Materials

The main workflow of the prototype application is analogous to the prototype application from Study 1 except for a few changes. In Study 2, applications in all groups included the newly designed Eggy as their AI agent. Eggy was redesigned to show more explicit feelings than the previous character. To make the application more realistic, food-diary features such as the history of food records and food image name tags were added to the main page (i) (see **Figure 9**). For both the explainable AI and monetary reward applications, typing in the correct answer for the wrong prediction was substituted with an interface that provided a list of possible answers in order to make participants experience more weariness, as shown in (ii). This was done because the participants in Study 1 were able to revise the result by simply selecting the correct answers in the list, and thus the process was too easy for them to feel any kind of load, leading to no difference among experimental groups. Given that Study 2 was designed to measure willingness to help objectively, users were required to experience weariness, so that only those who were more strongly motivated would actually provide more help. Therefore, the screen (ii) was displayed.



I guess I got it wrong.
Can you tell me the right answer?

Enter the right answer.
is the correct prediction.

Got it!
I'll go with this answer!

Food is recorded.
5 coin is paid.

Coke

Food is recorded.
10 coin is paid.

Alright, I'll keep this one as '?'.
Please give me the answer next time.

COIN HISTORY
+ 5 coin
1st Image | New photo input
+ 10 coin
2nd Image | New photo input + correction
+ 5 coin
3rd Image | New photo input
Total | 20 Coin

Figure 9. Interface of the application for monetary reward condition.

We made some changes to the material for the explainable AI condition. The application no longer included database related functions, which led to the exclusion of the point system. As the explainable AI condition no longer dealt with performance advancement, the application in Study 2 did not show the application's performance, explicit notification of performance advancement, or the actual elevation of prediction speed and accuracy. In total, 40 % of 150 pictures were predetermined to be incorrect. The application for the monetary reward condition did not explain the standard of the prediction outcome. Instead, it had a coin system that rewarded participants with coins whenever they took a picture for the application (5 coins) as shown in (iii), or made a correction (10 coins) as shown in (iv). If the user decided not to correct a wrong answer, (v) appeared, and the coins were not provided. Users were allowed to check their coin history through the coin wallet button on the main page (vi).

5.2.2. *Procedure*

Four prototype applications for the two by two experiment groups were developed with Android OS and they were installed on a test smartphone, a Galaxy Note 4. As in Study 1, in order to implement a Wizard of Oz experiment, an activated prototype application was presented to the participants. To make participants show their real intention in helping the AI app, the procedure in Study 2 was designed to require a greater cognitive load from each participant. We conducted a pre-test with two subjects for four sets of food images in order to identify the maximum number of images for people to process within 30 minutes. The pre-test results showed that none of the participants could take more than 110 pictures in 30 minutes. Consequently, a total of 150 pictures were prepared.

As in Study 1, all of the participants experienced exactly the same recognition performance. With the prototype application, the effects of explainable AI and monetary reward on willingness to help, with the mediating effect of trust, were tested. In Study 2, we objectively measured willingness to help with the actual helping behavior of the participants. Participants were told that they had a maximum of 30 minutes to use the application, and they could freely terminate usage whenever they desired. Additionally, they were told that everything was at their will. They were able to take pictures one by one, and to decide whether to correct the prediction or not. When they did not know the exact name of the food, they could do whatever they wanted. Participants earned a different number of coins according to their helping behavior. Once the session was over, the participants in the monetary reward condition received their participation reward in proportion to the number of coins earned.

5.2.3. *Measurements*

All of the questionnaires were measured on a 9-point Likert scale ranging from 1 (strongly disagree) to 9 (strongly agree). The following questionnaires for explainable AI and monetary reward were utilized for manipulation and hypotheses testing. Specific questionnaire items can be found in **Appendix C**.

Explainable AI

Unlike Study 1, the explanation condition was only composed of questions regarding explainable AI. The items measuring explanation were “The application I used sufficiently explained the standards of analysis.” and “I could understand the process of analysis made by the application I used.”

Monetary reward

The items regarding monetary reward asked participants about their perception of the benefit related to the monetary reward; “The application I used provided monetary reward for me.”, and “The monetary reward that I received is a benefit for me.”

Trust

Similar to study 1, the human-computer trust (HCT) scale was utilized to measure trust in participants. Questionnaires for cognition-based trust between human and computer were utilized again (Madsen & Gregor, 2000).

Willingness to help

Human beings' motives to help have been studied for a long time in the field of social science. Willingness to help was sometimes measured with questionnaires, but a variety of objective scales, which have resulted from observing human behavior, have also been utilized. Willingness to help has been measured by participants volunteering to reply to a student's request, picking up a paper for others (Isen & Levin, 1972), providing help to a slumped man in a doorway (Darley & Batson, 1973), and the numbers of tasks done for an experimenter (Berkowitz, 1987). Therefore, in order to overcome the limitations of the subjective scale used in Study 1, objective scales were used in Study 2.

Three measurements were used for the objective recording of participants' willingness to help: the number of pictures taken, the number of coins earned, and the number of neglecting errors. The number of pictures taken refers to the total number of pictures taken while using the application. The number of pictures taken was considered to represent participants' willingness to help in that people who are more willing to help were expected to take more pictures for the AI to learn more data. Moreover, the act of taking a picture requires participant's time and effort. Before the experiment was conducted, all participants were informed that they could terminate the experiment at any time. Since we increased the effort required to complete the experiment in Study 2 by increasing the number of tasks and making users type all of the corrections, the participants in Study 2 were much more willing to terminate the experiment within the 30-minute time limit. Thus, the participants more willing to provide help for the AI were expected to take more pictures than the ones who did not. The coins were considered to represent users' willingness to help because they accumulated 5 coins when a picture was taken, 10 coins when an error was corrected, and no coins when an error was neglected. The number of coins earned was not displayed in the condition without a monetary reward, but it was calculated in the back of the system. Lastly, the number of neglecting errors is the negative scale of the number of corrections. The negative scale of the number of corrections refers to the number of decisions to not make corrections. A negative value was used because when users were asked to fix the incorrect predictions, they could choose to correct the error or refuse to correct the error (see Figure 9(ii)). However, a considerable number of participants decided to correct all the predictions. Since making corrections was the users' default reaction, measuring the users' neglect was considered to reflect their willingness to help more effectively. However, as discussed in section 5.3.3, the number of neglecting errors did not meet the assumptions of normality and homogeneity of variance. Accordingly, the number of neglecting errors was analyzed with a non-parametric test and was not included in the PLS measurement.

5.2.4. Validity and Reliability

A total of 11 scales were developed; of these, 9 questions were developed from previous validated studies (Madsen & Gregor, 2000). Questions for the explainable AI and monetary reward were developed based on related studies (Gunning, 2017; Williamson, 1993). All of the questionnaires were translated into Korean by an expert with a degree in English Literature, then they were co-reviewed by three other experts majoring in HCI, and finally they were revised by an HCI expert. Verification tests conducted with Smart PLS 3.0 showed that the factor loadings of all the items were above 0.70. Cronbach's alphas and composite reliabilities for all of the factors exceeded 0.70, indicating the reliability of the research (**Appendix C**). In addition, discriminant validity was achieved for Study 2. Square roots of the average variance were extracted (diagonal elements in **Appendix D**) and were shown to be larger than the correlation between one latent variable and the other variables (off-diagonal elements).

5.3. Results

Participants were recruited from a variety of online communities similar to Study 1. A total of 75 participants (33 males; 42 females; average age of 24) completed our experiment. Participants were given a reward worth 8 USD once the experiment was completed. Meanwhile, participants of monetary reward condition received their reward ranged from 7 USD to 13 USD. All of the participants were educated to undergraduate or graduate school level. Out of the 75 participants recruited, 6 participants' data were eliminated. Two were eliminated due to their low attention during the test. Participants were identified to be too distracted to be included in the sample if they did not pass more than two of the manipulation questions, which included the name of the character and the standard of the image recognition. A box-plot was made using SPSS 23. Values that did not fall within the interquartile range multiplied by 1.5 were considered as outliers. Consequently, data from 69 participants (29 males; 40 females) were analyzed in the study. Nineteen participants in the control group, 17 in explainable AI condition, 16 in monetary reward condition, and 17 in both explainable AI and monetary reward conditions were analysed. No statistically significant difference was found between different gender groups.

5.3.1. Manipulation Check

Two manipulation tests were implemented where two independent variables, explainable AI and monetary reward, were manipulated well. The manipulation of explainable AI was measured using the explainable AI measures, and the two-tailed t-test result demonstrated that there was a significant difference (t(67) = 2.258, p= 0.027*) between the groups with (M =

5.93, SD = 1.65) and without explainable AI (M = 4.97, SD = 1.86). A t-test for the monetary reward condition was also conducted. There was a significant difference in perceived monetary reward (t(67) = 8.459, p = 0.000***) between the group with (M = 7.74, SD = 1.31) and without the monetary reward (M = 4.68, SD = 1.66). The manipulation check results are shown in **Table 4.**

Table 4. Manipulation check results of study 2

Independent variables	**Mean (std. dev.)**		**t-statistic**		
	High	**Low**	**Mean difference**	**T statistic**	**Significance level**
Explainable AI	5.93 (1.65)	4.97 (1.86)	0.955	t(67) = 2.258	p = 0.027*
Monetary Reward	7.74 (1.31)	4.68 (1.66)	3.062	t(67) = 8.459	p = 0.000***

*p < 0.05, **p < 0.01, ***p<0.001.

5.3.2. Hypotheses Testing (Structural Equation Model)

Structural equation modeling (Anderson & Gerbing, 1988; Bagozzi & Yi, 2012) was used in Study 2, as in Study 1, and was also bootstrapped with 200 resamples in order to calculate the t-statistics (Chin, 1998). **Figure 10** and **Table 5** show the entire path coefficients of the model and the associated statistical significance. Monetary reward (path coefficient = 0.216*) had a direct effect on willingness to help, but explainable AI did not (path coefficient = 0.219). In addition, explainable AI was shown to be positively associated with trust (path coefficient = 0.526***), but monetary reward did not (path coefficient = 0.060). Trust did not affect willingness to help (path coefficient = 0.100), resulting in an insignificant mediating effect.

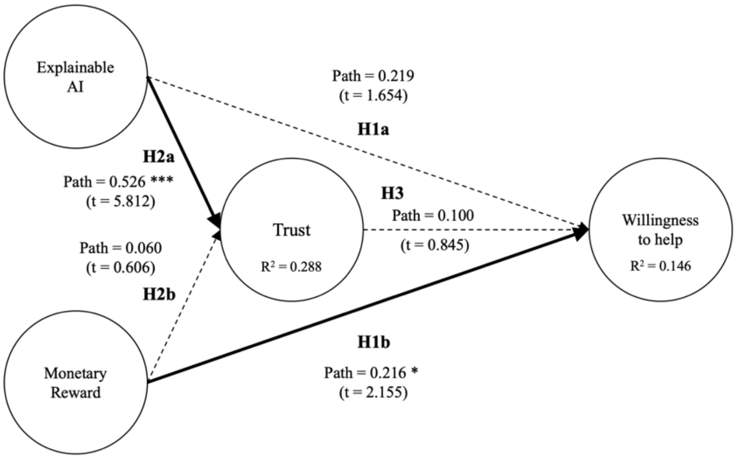


Figure 10. Structural Model of Study 2

Table 5. Partial least square results of study 2

Path (Operational Variables)	**Hypothesis**	**Path Coefficient**	**t-statistic**	**p-value**	**Hypotheses Supported**
Explainable AI → Willingness to help	H1a	0.219	1.654	0.100	-
Monetary Reward → Willingness to help	H1b	0.216	2.155	0.032*	supported
Explainable AI → Trust (HCT)	H2a	0.526	5.812	0.000***	supported
Monetary Reward → Trust (HCT)	H2b	0.060	0.606	0.545	-
Trust (HCT) → Human Help	H3	0.100	0.845	0.399	-

*p < 0.05, **p < 0.01, ***p<0.001.

5.3.3. *Additional Analysis with ANOVA and Kruskal-Wallis Test*

In order to take a closer look at the effects of explainable AI and monetary reward on willingness to help, a two-way ANOVA was implemented. Normality and variance homogeneity were tested by skewness/kurtosis and Levene's test. For the number of pictures taken and the number of coins earned, the values of skewness ranged between -2 and 2, and the value of kurtosis ranged between -7 to 7, indicating that the data met the normality assumption. The F-values of the Levene's test are 1.555 (p = 0.209; for number of pictures), and 3.898 (p = 0.013; for the number of coins). Because the p-values are larger than 0.01, the homogeneity of variance is met. On the other hand, the number of neglecting errors did not pass the normality and homogeneity test. This is because a number of participants decided to correct all the predictions. Only 34 participants decided to neglect error, indicating that less than half of the participants decided to neglect incorrect predictions. This was due to limitations of the experimental context. As a result of this problem, the distribution of decisions to neglect errors was not a normal distribution. Therefore, the Kruskal–Wallis test, which is a non-parametric test, was implemented in place of ANOVA.

The results of the number of pictures taken and the number of coins earned are displayed in **Table 6**. The results showed that monetary reward had a significant effect on the number of coins earned (F(1,65) = 4.802, p = 0.032). However, the monetary reward did not affect the number of pictures taken by participants in the different groups. In the post-interview, participants said that they had a goal for the number to conduct because they were participating in an experiment [P15, 74, 75]. Therefore, no significant difference between the groups was found in the number of iterations. In addition, explainable AI did not show any significant effect on the objective scales of willingness to help.

Table 6. ANOVA results of study 2

IV	**DV**	**Mean (std. dev.)**		**Hypotheses Testing**			**Hypotheses Supported**
		high	low	F statistic	Significant level	Eta square	
Explainable AI	Number of pictures	75.71 (16.12)	72.46 (17.33)	F(1,65) = 0.486	p = 0.488	η2 = 0.007	-
	Number of coins	533.68 (108.14)	499.86 (134.54)	F(1,65) = 1.067	p = 0.305	η2 = 0.016	-
Monetary	Number of	76.61	71.72	F(1,65) =	p = 0.244	η2 = 0.021	-

Reward	pictures	(16.71)	(19.91)	1.384			
	Number of coins	550.00 (79.62)	485.83 (146.12)	F(1,65) = 4.802	p = 0.032*	η2 = 0.069	supported

*p < 0.05, **p < 0.01, ***p<0.001.

In addition, there was a statistically significant difference between the condition with monetary reward (rank = 30.15) and without monetary reward (rank = 39.44) in the number of neglecting errors (H(1) = 4.245, p = 0.039). However, the explainable AI condition did not make a statistically significant difference in the number of neglecting errors (H(1) = 0.448, p = 0.503). The result of the Kruskal–Wallis test can be found in **Table 4**.

Table 4. Kruskal-Wallis test result of neglecting errors

	Chi-square	**df**	**Asymp. Sig**
Explainable AI	0.448	1	0.503
Monetary Reward	4.245	1	0.039*

*p < 0.05, **p < 0.01, ***p<0.001.

5.4. Discussion

In Study 2, willingness to help was measured by the number of pictures taken, the number of coins earned, and the number of neglecting errors, in contrast to Study 1 that measured willingness to help using a subjective scale. PLS showed that monetary reward had a direct effect on willingness to help. For the trust construct, although it did not have a mediating effect on willingness to help, explainable AI was shown to have a strong positive correlation regarding trust in the AI.

The insignificance of the relationship between monetary reward and trust (H2b) and the relationship between trust and willingness to help (H3) seem to have resulted from the direct effect of monetary reward on willingness to help. That is, the insignificant indirect effect of monetary reward on willingness to help (H2b-H3) resulted because the direct effect of monetary reward on willingness to help was too strong. However, in Study 1, despite the significant direct effect of empathy on willingness to help, H3 was supported. Consequently, it can be said that monetary reward is such a powerful incentive that it causes users to be willing to help but neutralizes the effects of other variables. This argument is the basis for the strong, significant results of the ANOVA for monetary reward and the insignificant results of the PLS for explainable AI and willingness to help (H1a). In summary, explainable AI is not a stimulus as strong as monetary reward to engage users in the loop, but it does have a strong positive effect on the trust in the AI. Moreover, the ANOVA results did not reveal any differences between the groups with and without explainable AI. It is possible that because Eggy, the AI agent that triggers empathy, was included in all of the prototypes, it might have affected users' responses. The AI agent not only showed emotion but also conveyed some simple conversations such as the fact that it was processing the picture, or that it was not easy to find the answer. Thus, a certain level of explanation may have been perceived by the group without explainable AI, resulting in no significant difference.

In Study 2, participants could fix the incorrect predictions of the AI, but they also could choose to neglect the errors. The ANOVA results showed that the difference in the number of coins earned was statistically significant, but the number of pictures taken was not. These results show that participants receiving monetary rewards made more corrections than participants not receiving monetary rewards. While participants of the monetary reward condition tried to adjust the AI predictions as much as possible, participants without the monetary reward did not put in similar effort. The records in the application revealed that when participants in the monetary reward group faced an exotic food, they fixed the result with a similar-looking food with which they were more familiar [P21, 27, 29, 50, 72]. Moreover, participants [29, 37, 47] in this group showed that they tried to adjust food names that were not specific, even when the prediction was correct. Participants without monetary reward were prone to neglecting errors. Interestingly, when they were asked why they did so, 13 participants said that it was a mistake, or because they did not recognize the food. This indicates that when it comes to monetary reward, people do not respond honestly and that cheating to obtain a higher reward is much more likely to occur.

Moreover, the Kruskal–Wallis test result for neglecting errors shows that when a monetary reward is not provided, people are more likely to decide not to correct the wrong predictions of the AI, even when social pressure is exerted by the experimental environment. In other words, providing a monetary reward reduces the possibility of not correcting the AI's incorrect predictions. This result supports the implication of this study, which suggests that monetary reward is the most robust means for increasing willingness to help. In summary, a monetary reward is an effective source of human help; however, it can also be an incentive for dishonest behavior. Therefore, in order to utilize monetary rewards in an AI application as a reward for users' helpful behavior, sophisticated design is required to avoid moral temptations.

6. General Discussion

Human intervention in the HITL system is crucial for AI to progress interactively. However, studies regarding interactions designed to induce human intervention have received little attention. The aim of this study was to find critical factors that encourage human participants to engage in the loop with AI, specifically to correct inaccurate AI predictions. Study 1 adopted a social science theory of motivation to help, including egoistic and altruistic motivations, to design stimuli that would induce the users' willingness to help. An explanation of the AI and an explanation of how the process would improve the performance of the AI were designed to provide utilitarian benefits and therefore trigger an egoistic motivation. In addition, a humanlike AI agent called Eggy was designed to induce empathy, an altruistic motivation, in users. Finally, trust in the AI system was posited as a mediating factor between the two motivations and willingness to help. The results revealed that utilitarian benefit has an indirect effect on willingness to help whereas empathy has a direct effect on willingness to help.

There were some limitations to Study 1. The utilitarian benefit condition contained two different constructs, explainable AI and benefit, and the participants also tried to give positive answers for the experimenters, leading to an insignificant difference between the groups. Thus, in Study 2 the utilitarian benefit condition was separated into the explainable AI and monetary reward conditions. In this case, the benefits were provided in the form of monetary rewards in order to provide a stronger benefit stimulus. Willingness to help was also measured objectively based on three different components of helping behavior: the number of pictures taken, the number of coins earned, and the number of neglecting errors. In order to objectively measure participants' willingness to help, the experimental design was adjusted to make the process more tedious. The application presented 150 food images to each participant but this volume made it impossible for the participant to complete the application within the allotted time. Furthermore, corrections had to be made by typing every single letter. For the monetary reward condition, coins were accumulated in the user's application according to their behavior: 0 for declining to help, 5 for taking a picture, and 10 for correcting an error. Using PLS, Study 2 revealed that monetary reward had a direct effect on willingness to help and explainable AI was significantly associated with trust.

In conclusion, the two studies demonstrated that empathy and monetary reward positively affected users' willingness to help. The results of Study 1 indicate that the utilitarian benefit construct did not have a direct effect on willingness to help. Since the utilitarian benefit construct in Study 1 contains both explanation and benefit related aspects, the results of Study 2, composed of explainable AI and monetary reward, can explain the insignificance of the utilitarian benefit on willingness to help. The two constructs of Study 2, explainable AI and monetary reward, were developed based on the explanation related aspect and the benefit related aspect of the utilitarian benefit construct, respectively. In Study 2, explainable AI did not have a direct effect on willingness to help but monetary reward did. On the other hand, the monetary reward, which encourages egoistic motivation, was shown to be a robust trigger for inducing users' willingness to help. Therefore, providing a monetary reward can be a very effective incentive to induce human intervention. However, the monetary reward was so powerful that some of the participants showed cheating behavior. According to the post-hoc analysis of system log data, some participants fixed results that did not require any adjustment. Thus, further studies should be conducted to resolve the adverse effects of monetary reward in HITL. Explanation had a greater impact on trust than on help. This was consistent across both Study 1 and 2. Even though the effect of explanation was not as strong as that of the monetary reward on willingness to help, it was significantly effective in forming trust.

This study makes both theoretical and practical contributions. Theoretically, this study applies social studies of help motivation to the HCI field in order to provide a positive user experience with an imperfect AI. Benefits, based on egoistic motivation, and empathy, based on altruistic motivation, were originally shown to be effective when inducing help from another human being. This study has major significance in that these components were also shown to be effective in inducing help for AI. Previously, the majority of studies that have measured participants' willingness to help have utilized a subjective scale. Our first study revealed that a subjective scale of willingness to help shows low validity due to the experimental context; it also shows that measuring willingness to help with self-reporting questionnaires is inadequate, and so we developed objective scales of willingness to help. This study contributes to the field of HCI by developing the components of objective measurements and significantly validating these measures.

The practical implications of this study include interaction components that can induce users' voluntary engagement in the loop with AI. With the development of AI, there will be an increasing number of AI services. Currently, the accuracy level of AI is often considered to be low. Thus, interactions that enhance prediction accuracy and improve the user's experience with AI are needed. This study confirms that empathy and monetary reward can encourage users to help AI. Moreover, explainable AI was shown to increase trust in AI. These components can be added to existing AI services to enhance their accuracy and trust. Interactive designs that induce users to help are useful because AI accuracy can be interactively and effectively enhanced based on users' corrections for incorrect predictions. Furthermore, enhanced trust in AI resulting from explainable AI and monetary reward can lead to a more positive experience with AI services.

This study is limited because the experiment was not implemented in situ or over a long term. In order to study users' actual intention to help, long-term field research with actual working AI should be conducted as willingness to help can change over time. Further research should also be done to investigate actual human intention with real users of an AI application. The current AI is not perfect, so it is important to consider utilitarian benefit, empathy, explainable AI, and monetary reward during the design process for effective interactions with AI, to build trust, and to encourage users to provide help.

Reference

Amos, O. M. (1982). Empirical analysis of motives underlying individual contributions to charity. *Atlantic Economic Journal, 10*(4), 45-52.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological bulletin, 103*(3), 411.

Arnold, V., Clark, N., Collier, P. A., Leech, S. A., & Sutton, S. G. (2004). Explanation provision and use in an intelligent decision aid. *Intelligent Systems in Accounting, Finance & Management: International Journal, 12*(1), 5-27.

Ba, S., & Pavlou, P. A. (2002). Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior. *MIS quarterly*, 243-268.

Ba, S., Whinston, A. B., & Zhang, H. (2003). Building trust in online auction markets through an economic incentive mechanism. *Decision Support Systems, 35*(3), 273-286.

Babin, B. J., Darden, W. R., & Griffin, M. (1994). Work and/or fun: measuring hedonic and utilitarian shopping value. *Journal of consumer research, 20*(4), 644-656.

Bagozzi, R. P., & Yi, Y. (2012). Specification, evaluation, and interpretation of structural equation models. *Journal of the academy of marketing science, 40*(1), 8-34.

Batson, C. D. (1987). Prosocial motivation: Is it ever truly altruistic? *Advances in experimental social psychology* (Vol. 20, pp. 65-122): Elsevier.

Batson, C. D., Chang, J., Orr, R., & Rowland, J. (2002). Empathy, attitudes, and action: Can feeling for a member of a stigmatized group motivate one to help the group? *Personality and Social Psychology Bulletin, 28*(12), 1656-1666.

Batson, C. D., Duncan, B. D., Ackerman, P., Buckley, T., & Birch, K. (1981). Is empathic emotion a source of altruistic motivation? *Journal of personality and social psychology, 40*(2), 290.

Batson, C. D., O'Quin, K., Fultz, J., Vanderplas, M., & Isen, A. M. (1983). Influence of self-reported distress and empathy on egoistic versus altruistic motivation to help. *Journal of personality and social psychology, 45*(3), 706.

Becker, M. H., & Maiman, L. A. (1975). Sociobehavioral determinants of compliance with health and medical care recommendations. *Medical care*, 10-24.

Berkowitz, L. (1987). Mood, self-awareness, and willingness to help. *Journal of personality and social psychology, 52*(4), 721.

Bezdek, J. C., Keller, J., Krisnapuram, R., & Pal, N. (1999). *Fuzzy models and algorithms for pattern recognition and image processing* (Vol. 4): Springer Science & Business Media.

Brown, D., & Grinter, R. E. (2016). *Designing for transient use: A human-in-the-loop translation platform for refugees.* Paper presented at the Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.

Callison-Burch, C. (2009). *Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk.* Paper presented at the Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1.

Chandon, P., Wansink, B., & Laurent, G. (2000). A benefit congruency framework of sales promotion effectiveness. *Journal of marketing, 64*(4), 65-81.

Chen, G.-D., Lee, J.-H., Wang, C.-Y., Chao, P.-Y., Li, L.-Y., & Lee, T.-Y. (2012). An empathic avatar in a computer-aided learning program to encourage and persuade learners. *Journal of Educational Technology & Society, 15*(2), 62-72.

Chin, W. W. (1998). The partial least squares approach to structural equation modeling. *Modern methods for business research, 295*(2), 295-336.

Chung, J. B., & Kim, H.-K. (2009). Competition, economic benefits, trust, and risk perception in siting a potentially hazardous facility. *Landscape and Urban Planning, 91*(1), 8-16.

Coulter, K. S., & Coulter, R. A. (2002). Determinants of trust in a service provider: the moderating role of length of relationship. *Journal of services marketing, 16*(1), 35-50.

Darley, J. M., & Batson, C. D. (1973). " From Jerusalem to Jericho": A study of situational and dispositional variables in helping behavior. *Journal of personality and social psychology, 27*(1), 100.

DEFENSE, U. D. O. (1994). Directive 5000.59 ‐ M: DoD M&S Glossary.

Delgado, M. R. (2007). Reward ‐ related responses in the human striatum. *Annals of the New York Academy of Sciences, 1104*(1), 70-88.

Dill, K., & Rabin, S. (2013). What is game AI? *Game AI pro: Collected wisdom of game AI professionals*, 3-10.

Duflo, E., & Saez, E. (2003). The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment. *The Quarterly journal of economics, 118*(3), 815-842.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International journal of human-computer studies, 58*(6), 697-718.

Eisenberg, N., & Miller, P. A. (1987). The relation of empathy to prosocial and related behaviors. *Psychological bulletin, 101*(1), 91.

Fails, J. A., & Olsen Jr, D. R. (2003). *Interactive machine learning.* Paper presented at the Proceedings of the 8th international conference on Intelligent user interfaces.

Fehr, E., & Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *Journal of economic perspectives, 14*(3), 159-181.

Feng, J., Lazar, J., & Preece, J. (2004). Empathy and online interpersonal trust: A fragile relationship. *Behaviour & Information Technology, 23*(2), 97-

106.

Gounaris, S. P. (2005). Trust and commitment influences on customer retention: insights from business-to-business services. *Journal of Business research, 58*(2), 126-140.

Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*.

Hair Jr, J. F., Hult, G. T. M., Ringle, C., & Sarstedt, M. (2016). *A primer on partial least squares structural equation modeling (PLS-SEM)*: Sage Publications.

Hoff, S., Jaffurs, P., Enriquez, M., & Wilde, Q. (2018). Snap-n-Snack: a Food Image Recognition Application.

Hojat, M., Louis, D. Z., Maxwell, K., Markham, F., Wender, R., & Gonnella, J. S. (2010). Patient perceptions of physician empathy, satisfaction with physician, interpersonal trust, and compliance. *International Journal of Medical Education, 1*, 83.

Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics, 3*(2), 119-131.

Imran, M., Castillo, C., Lucas, J., Meier, P., & Vieweg, S. (2014). *AIDR: Artificial intelligence for disaster response.* Paper presented at the Proceedings of the 23rd International Conference on World Wide Web.

Isen, A. M., & Levin, P. F. (1972). Effect of feeling good on helping: cookies and kindness. *Journal of personality and social psychology, 21*(3), 384.

Johnson, E., Hervás, R., Gutiérrez López de la Franca, C., Mondéjar, T., Ochoa, S. F., & Favela, J. (2018). Assessing empathy and managing emotions through interactions with an affective avatar. *Health informatics journal, 24*(2), 182-193.

Kim, S. S., Kaplowitz, S., & Johnston, M. V. (2004). The effects of physician empathy on patient satisfaction and compliance. *Evaluation & the Health Professions, 27*(3), 237-251.

Klein, G., Moon, B., & Hoffman, R. R. (2006). Making sense of sensemaking 1: Alternative perspectives. *IEEE intelligent systems*(4), 70-73.

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature, 435*(7042), 673.

Koster, F. (2007). Globalization, social structure, and the willingness to help others: a multilevel analysis across 26 countries. *European Sociological Review, 23*(4), 537-551.

Kulesza, T., Burnett, M., Wong, W.-K., & Stumpf, S. (2015). *Principles of explanatory debugging to personalize interactive machine learning.* Paper presented at the Proceedings of the 20th international conference on intelligent user interfaces.

Lasecki, W. S., Wesley, R., Nichols, J., Kulkarni, A., Allen, J. F., & Bigham, J. P. (2013). *Chorus: a crowd-powered conversational assistant.* Paper presented at the Proceedings of the 26th annual ACM symposium on User interface software and technology.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*(10), 1243-1270.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors, 46*(1), 50-80.

Leeper, A. E., Hsiao, K., Ciocarlie, M., Takayama, L., & Gossow, D. (2012). *Strategies for human-in-the-loop robotic grasping.* Paper presented at the Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction.

Libera, C. D., & Chelazzi, L. (2006). Visual selective attention and the effects of monetary rewards. *Psychological science, 17*(3), 222-227.

Lim, B. Y., & Dey, A. K. (2011). *Design of an intelligible mobile context-aware application.* Paper presented at the Proceedings of the 13th international conference on human computer interaction with mobile devices and services.

Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). *Why and why not explanations improve the intelligibility of context-aware intelligent systems.* Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Liu, C., Cao, Y., Luo, Y., Chen, G., Vokkarane, V., & Ma, Y. (2016). *Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment.* Paper presented at the International Conference on Smart Homes and Health Telematics.

Liu, Z., Qiao, F., Long, H., & Li, G. (2018). *GazeLabel: A Cost-free Data Labeling System with Public Displays using Eye-tracking.* Paper presented at the Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems.

Madsen, M., & Gregor, S. (2000). *Measuring human-computer trust.* Paper presented at the 11th australasian conference on information systems.

Martınez-Miranda, J., & Aldea, A. (2005). Emotions in human and artificial intelligence. *Computers in Human Behavior, 21*(2), 323-341.

McAllister, D. J. (1995). Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of management journal, 38*(1), 24-59.

Meltzoff, A. N., & Moore, M. K. (1983). Newborn infants imitate adult facial gestures. *Child development*, 702-709.

Mercado, J. E., Rupp, M. A., Chen, J. Y., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent agent transparency in human–agent teaming for Multi-UxV management. *Human factors, 58*(3), 401-415.

Meyers, A., Johnston, N., Rathod, V., Korattikara, A., Gorban, A., Silberman, N., . . . Murphy, K. P. (2015). *Im2Calories: towards an automated mobile vision food diary.* Paper presented at the Proceedings of the IEEE International Conference on Computer Vision.

Miller, T. (2017). Explanation in artificial intelligence: insights from the social sciences. *arXiv preprint arXiv:1706.07269.*

Millon, T., Lerner, M. J., & Weiner, I. B. (2003). *Handbook of Psychology: Volume 5, Personality and Social Psychology*: New Jersey: John Wiley & Sons, Inc.

Misselhorn, C. (2009). Empathy with inanimate objects and the uncanny valley. *Minds and Machines, 19*(3), 345.

Mori, M. (1970). The uncanny valley. *Energy, 7*(4), 33-35.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies, 27*(5-6), 527-539.

Naor, M. (1996). Verification of a human in the loop or Identification via the Turing Test. *Unpublished draft from http://www. wisdom. weizmann. ac. il/~*

naor/PAPERS/human abs. html.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of social issues, 56*(1), 81-103.

O'Leary, D. E. (2013). Artificial intelligence and big data. *IEEE intelligent systems, 28*(2), 96-99.

O'Brien, H. L. (2010). The influence of hedonic and utilitarian motivations on user engagement: The case of online shopping experiences. *Interacting with computers, 22*(5), 344-352.

Paiva, A., Dias, J., Sobral, D., Aylett, R., Woods, S., Hall, L., & Zoll, C. (2005). Learning by feeling: Evoking empathy with synthetic characters. *Applied Artificial Intelligence, 19*(3-4), 235-266.

Pallas, F., Mittal, V., & Groening, C. (2014). Allocation of resources to customer satisfaction and delight based on utilitarian and hedonic benefits.

Pearl, C. (2016). *Designing Voice User Interfaces: Principles of Conversational Experiences*: " O'Reilly Media, Inc.".

Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology, 20*(1), 5-14.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *Why should i trust you?: Explaining the predictions of any classifier.* Paper presented at the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.

Ringle, C. M., Wende, S., & Becker, J.-M. (2015). SmartPLS 3. *Boenningstedt: SmartPLS GmbH, http://www. smartpls. com*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., . . . Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision, 115*(3), 211-252. doi:10.1007/s11263-015-0816-y

Shin, D. (2018). Empathy and embodied experience in virtual environment: To what extent can virtual reality stimulate empathy and embodied experience? *Computers in Human Behavior, 78*, 64-73.

Shin, D., & Biocca, F. (2018). Exploring immersive experience in journalism. *New media & society, 20*(8), 2800-2823.

Shin, D.-H. (2010). The effects of trust, security and privacy in social networking: A security-based approach to understand the pattern of adoption. *Interacting with computers, 22*(5), 428-438.

Shin, D.-H. (2013). User experience in social commerce: in friends we trust. *Behaviour & Information Technology, 32*(1), 52-67.

Shoham, Y., Perrault, R., Brynjolfsson, E., & Clark, J. (2017). Artificial Intelligence Index—2017 Annual Report.

Simon, H. A. (1993). Altruism and economics. *The American Economic Review, 83*(2), 156-161.

Spaulding, D. G., & Plank, R. E. (2007). Selling automobiles at retail: is empathy important? *Marketing Management Journal, 17*(2).

Twenge, J. M., Baumeister, R. F., DeWall, C. N., Ciarocco, N. J., & Bartels, J. M. (2007). Social exclusion decreases prosocial behavior. *Journal of personality and social psychology, 92*(1), 56.

Vallerand, R. J. (1997). Toward a hierarchical model of intrinsic and extrinsic motivation *Advances in experimental social psychology* (Vol. 29, pp. 271-360): Elsevier.

Weiner, B. (1980). A cognitive (attribution)-emotion-action model of motivated behavior: An analysis of judgments of help-giving. *Journal of personality and social psychology, 39*(2), 186.

Williamson, O. E. (1993). Calculativeness, trust, and economic organization. *The journal of law and economics, 36*(1, Part 2), 453-486.

Yang, A., Hsee, C., & Urminsky, O. (2014). Eager to help yet reluctant to give: How pro-social effort and pro-social choices diverge.

Zanzotto, F. M. (2017). Human-in-the-loop Artificial Intelligence. *arXiv preprint arXiv:1710.08191*.

Appendix A. Convergent Validity and Reliability of Study 1

Construct	Measurement Items	Factor Loading	AVE	Composite Reliability	Cronbach's Alpha
Utilitarian benefit 1	The application provided benefit for me.	0.648	0.663	0.853	0.739
Utilitarian benefit 2	The performance of the application has improved.	0.861			
Utilitarian benefit 3	The performance improvement of the application met my expectations.	0.910			
Empathy 1	I was moved when I was using the application.	0.776	0.728	0.914	0.884
Empathy 2	I felt compassionate when I was using the application.	0.856			
Empathy 3	I felt soft-hearted when I was using the application.	0.876			
Empathy 4	I felt tender when I was using the application.	0.899			
Trust 1	The application performs reliably.	0.820	0.681	0.914	0.883
Trust 2	I can rely on the application to function properly.	0.832			
Trust 3	The application analyzes problems consistently.	0.785			
Trust 4	The application uses appropriate methods to reach decisions.	0.852			
Trust 5	The application has sound knowledge of this type of problem built into it.	0.837			
Willingness to help 1	When the application asked me to correct the food recognition result, I was willing to help it.	0.924	0.805	0.892	0.762
Willingness to help 2	When the application asked me to correct the food recognition result, I was annoyed. (reverse-coded)	0.869			

Appendix B. Discriminant Validity of Study 1

Construct	Explanation	Empathy	Trust	Willingness to help
Utilitarian benefit	**(0.814)**			
Empathy	0.042	**(0.853)**		
Trust	0.626	-0.047	**(0.825)**	
Willingness to help	0.255	0.335	0.308	**(0.897)**

Appendix C. Convergent Validity and Reliability of Study 2

Construct	Measurement Items	Factor Loading	AVE	Composite Reliability	Cronbach's Alpha
Explainable AI 1	The application I used sufficiently explained the standards of analysis.	0.939	0.817	0.899	0.783
Explainable AI 2	I could understand the process of the analysis made by the application I used.	0.867			
Monetary reward 1	The application I used provided monetary reward for me.	0.961	0.931	0.965	0.927
Monetary reward 2	The monetary reward that I received is a benefit for me.	0.969			
Trust 1	The application performs reliably.	0.848	0.677	0.913	0.881
Trust 2	I can rely on the application to function properly.	0.835			
Trust 3	The application analyzes problems consistently.	0.791			
Trust 4	The application uses appropriate methods to reach	0.794			

	decisions.				
Trust 5	The application has sound knowledge of this type of problem built into it.	0.843			
Willingness to help 1	Objective Scale – the number of coins earned	0.982	0.967	0.983	0.965
Willingness to help 2	Objective Scale – the number of pictures taken	0.984			

Appendix D. Discriminant Validity of Study 2

Construct	Explainable AI	Monetary reward	Trust	Willingness to help
Explainable AI	**(0.904)**			
Monetary reward	0.132	**(0.965)**		
Trust	0.534	0.129	**(0.823)**	
Willingness to help	0.3	0.258	0.244	**(0.983)**

Appendix E. Partial least square results of Indirect effects in Study 1

Path (Operational Variables)	**original sample (O)**	**sample mean (M)**	**standard deviation**	**t-statistic**	**p-value**
Utilitarian Benefit → Trust → Willingness to help	0.180	0.172	0.085	2.107	0.036 *
Empathy → Trust → Willingness to help	-0.021	-0.023	0.038	0.556	0.579

*p < 0.05, **p < 0.01, ***p<0.001.

[Research Highlights]

- Utilitarian benefit indirectly increases willingness to help, mediated by trust
- Empathy directly increases willingness to help
- Explainable AI and monetary reward increase willingness to help
- Explainable AI has a greater impact on trust than on willingness to help
- Monetary reward is shown to be the most robust trigger of willingness to help