Human Computation (2022) 9:1:1-21

ISSN: 2330-8001, DOI: 10.15346/hc.v9i1.134

AI-TAM: a model to investigate user acceptance and collaborative intention in *human-in-the-loop* AI applications

ILARIA BARONI, CEFRIEL - POLITECNICO DI MILANO

GLORIA RE CALEGARI, CEFRIEL - POLITECNICO DI MILANO

DAMIANO SCANDOLARI, CEFRIEL - POLITECNICO DI MILANO

IRENE CELINO, CEFRIEL - POLITECNICO DI MILANO

ABSTRACT

More and more frequently, digital applications make use of Artificial Intelligence (AI) capabilities to provide advanced features; on the other hand, human-in-the-loop approaches are on the rise to involve people in AI-powered pipelines for data collection, results validation and decision-making. Does the introduction of AI features affect user acceptance? Does the AI result quality affect people's willingness to use such applications? Does the additional user effort required in human-in-the-loop mechanisms change the application adoption and use?

This study aims to provide a reference approach to answer those questions. We propose a model that extends the Technology Acceptance Model (TAM) with further constructs explicitly related to AI – user trust in AI and perceived quality of AI output, from explainable AI (XAI) literature – and collaborative intention – willingness to contribute to AI pipelines.

We tested the proposed model with an application for car damage claim reporting with AI-powered damage estimation for insurance customers. The results showed that the XAI related factors have a strong and positive effect on behavioral intention, perceived usefulness, and ease of use of the application. Moreover, there is a strong link between behavioral intention and collaborative intention, indicating that indeed human-in-the-loop approaches can be successfully adopted in final user applications.

1. INTRODUCTION

Artificial Intelligence (AI) permeates many aspects of our (digital) life and we are more and more used to consider machine processing an indispensable ingredient for our daily tasks: we ask virtual

assistants to support our home automation and we rely on spam filters to free us from undesired emails. In some cases, we are also tolerant to machine mistakes, like when we have to repeat a command to our virtual assistant or we find unwanted messages in our inbox, because we still have the notion that AI cannot perfectly replace a human intelligence, and that a machine, just like people, can make mistakes, even if they may be of a different kind.

The relationship between humans and AI is therefore an important and very interesting topic of recent research, especially because human-machine cooperation pipelines are more and more frequently adopted to make the best out of both worlds: *human-in-the-loop* is the paradigm to intertwine human and machine intelligence, exploiting the different capabilities in a virtuous cycle in which not only the AI supports people, but also people offer their human skills to help machines. Some examples of the latter case are collection of training data for machine learning through crowdsourcing (Howe, 2008), verification or cross-validation of machine results through human computation (Law and Ahn, 2011), or incremental and iterative improvement of algorithmic performance through active learning (Settles, 2009) or interactive machine learning (Amershi et al., 2012).

In all cases of cooperation between humans and AI, trust plays a very important role: the algorithm must be "inherently" trustworthy, in that it does not introduce biases or unethical behaviours, and the users should have at least some level of "personal" trust in the AI, in that they are confident that the machine provides some gain or advantage in executing their tasks.

In this paper, we aim at assessing the different aspects that characterize the relationship between users and an AI-powered system adopting a human-in-the-loop paradigm. The original contribution of this paper is the definition of a measurement of the users' perception of AI, their willingness to actively contribute in a human-in-the-loop mechanism, and the interplay between those factors and their acceptance and adoption of this system.

The remainder of this paper is organized as follows: we first present the theoretical bases on which this study is grounded in the *related work* section. In the *AI-TAM* section we then introduce our proposed user acceptance model with our research hypotheses. We describe the AI based application that we used to collect data and validate our model in the *scenario* section, followed by the details on the experimental protocol in the *methodology* section. We illustrate our study results in two sections: in the first one we offer the *preliminary analysis*, with the exploratory investigation of the scales, items reliability, and correlations among the variables; in the following section, we present the *results* from the confirmatory factor analysis and the final validated model. Finally, give an interpretation of the results in the *discussion* section and we close the paper with *conclusions and next steps*.

2. RELATED WORK

Trustworthy AI is a rising topic, not only in research, but also on mainstream media as well as among policy makers. Ethical guidelines for the adoption of AI solutions have been defined (AI HLEG, 2019) and legislation is being introduced to regulate the use of AI, especially in high-risk situations (European Commission, 2021), even if the ethical and legal aspect of AI involvement in decision making is still an open point (Green and Chen, 2019).

One of the main requirements for trustworthy AI is human agency and oversight: "AI systems should support human autonomy and decision-making, as prescribed by the principle of respect for

human autonomy" (AI HLEG, 2019). While AI can be adopted in different steps of the decision making process, human-in-the-loop specifically refers to the capability for human intervention in every decision cycle of an AI-powered system. A framework to help investigate whether an AI system being developed, deployed, procured, or used, adheres to the requirements of Trustworthy Artificial Intelligence has been recently introduced (AI HLEG, 2020), with a list of qualitative assessment questions for self-evaluation.

Several frameworks are indeed emerging to assess trustworthy AI, which covers qualitative and quantitative evaluation criteria throughout the AI system lifecycle. Some examples are the Trustworthy AI Implementation Framework (TAII) (Baker-Brunnbauer, 2009), Z-Inspection (Zicari et al., 2021), Values Criteria Indicators Observables (VCIO) (Hallensleben et al., 2009), and the data-driven research framework for trustworthy artificial intelligence (DaRe4TAI) (Thiebes et al., 2020).

The interplay of humans and machines is a delicate and multi-faceted topic, not easy to grasp and interpret. In (Hidalgo et al., 2021), A/B testing was applied to several scenarios to understand "how humans judge machines": for example, when evaluating an accident caused by a driver vs. a driverless car, it was noted how people judge humans by their intentions and machines by their outcomes. In some contexts, people are unwilling to use a machine even if it is more performant: in (Longoni et al., 2019) people declared to prefer a human radiologist even if an AI-powered system executing the same task was proved to be more precise, while in (Dietvorst et al., 2015), participants trusted humansâĂŹ feedback more than AI, even when provided with evidence that the AI performs better. When AI technology fails user perception, a deterioration of its performance may be perceived, an example being when failing to provide useful explanations (Ray et al., 2019), or when chatbots reveal a clear mismatch between the user expectations and the actual interaction experience (Luger and Sellen, 2016).

User understanding of AI is a trending topic in AI research as "eXplainable Artificial Intelligence" (XAI). Most literature focus on generating explanations out of black-box models, attempting to support the user in "objectively" understanding the internal functioning of the algorithms, adopting a scientific approach, and possibly generating a law/theory, more similarly to white-box models (Ribeiro et al., 2016; Arya et al., 2019). However, an explanation could also be intended in a "subjective" way, focusing on what the human user wants to know about the AI (Mittelstadt et al., 2019). To facilitate the process of understanding the AI, there is a need for effective, unbiased, and user-friendly explanations, which also allows the users to judge the fairness of the decisions taken by the AI (Dodge et al., 2019), even if fairness evaluation is itself influenced by many other factors (Wang et al., 2020).

Indeed, explanations take a primary role in the perception of AI systems, as described in (Miller, 2019), which also argue that the design of explainable artificial intelligence should rely on multidisciplinary domains including philosophy, cognitive psychology/science, and social psychology. The explainability of the AI (system and results) is strongly dependent on the usersâĂŹ mental model, which in cognitive psychology is the representation of how a person understands events and processes (Klein and Hoffman, 2008). Studying the usersâĂŹ mental model in XAI context means studying the userâĂŹs "subjective" understanding of the AI.

During the analysis of the perception of an AI system from a user's perspective, the elicitation methods to understand the user mental models are crucial. Many examples of such methods are

available in the literature, like the Think-Aloud Problem Solving Task (Williams et al. (Williams et al., 1983)), where participants think aloud during task execution, or Task Reflection of Retrospection Task (Fryer, 1939), in which participants are asked to describe their reasoning after a task execution. In the AI system, the understanding of the mental model is essential not only during an evaluation phase but also to establish successful cooperation between the AI and humans. The AI needs to prompt the appropriate mental triggers quickly (e.g., curiosity, which is another aspect to be analyzed in the human-AI interaction context), and the results of the feedback provided by the human needs to be categorized and analyzed to feed the categorization of the usersâĂŹ mental model. This aspect needs to be balanced with a good response of the system, because as described in the work from Honeycutt et al. (Honeycutt et al., 2020), the human-in-the-loop feedback lowered participantsâĂŹ trust in the AI and the perception of the accuracy, especially when it is not taken into consideration.

In (Hoffman et al., 2019) the authors discuss potential evaluation metrics for XAI systems. The work focused on different aspects: the goodness of the explanations (with respect to the user's prior knowledge and current goal), whether users are satisfied by explanations, how well they understand the AI system, how curiosity motivates the search for explanations, whether the userâĂŹs trust and reliance on the AI are appropriate, and the degree of success/effectiveness of the human-machine system. In particular, they investigate the problem of trust in XAI systems. End-users might have positive or negative biases toward AI systems, as well as they might trust its decisions and support in some contexts or for some specific goals while rejecting it completely in others. Trust in AI/automation systems is widely discussed in the literature, as well as the concept of trust itself, but not all the scales are applicable to XAI because being too tied to the specific application performances, or too specific of an application domain (e.g., acceptance of embodied agents in human-robotic interactions), or too unbalanced toward the trust on the application rather than on the AI contribution to the results. Hoffman thus made a comprehensive review and selected a reduced number of scales, which were mostly overlapping in their main items. In particular, the result of the analysis was the production of a Trust Scale Recommended for XAI, which investigates both the trust of the user on the XAI results and its perceived quality of the results provided. Most of the items of the scale are derived from the Cahour-Fourzy Scale (Cahour and Forzy, 2009) (with some rephrases inspired by other analyzed scales), with some items taken from (Jian et al., 2000), the Schaefer Scale (K.E., 2013), and adaptations from the Madsen-Gregor Scale (Madsen and Gregor, 2000).

3. AI-TAM MODEL

In this paper, we aim at defining a user acceptance model for AI systems adopting a human-in-the-loop mechanism and we propose a set of items to quantitatively evaluate the different factors influencing user acceptance.

To achieve this objective, we started from the Technology Acceptance Model (TAM) (Davis, 1989), a model adopted from the Theory of Reasoned Action (TRA) (Fishbein and Ajzen, 1977). TAM is frequently used to predict and analyze the technology reception and adoption from users, it bases its analysis on the testers' *perceived usefulness* and *ease of use*, and how these two main constructs affect the *behavioral intention* to use the system. As theorized in the TAM, the perceived usefulness is the individualâĂŹs silent belief that using a particular technology will improve his/her performance,

the perceived ease of use is the individualâĂŹs silent belief that using a particular technology is free of effort, and the behavioral intention is the individualâĂŹs perceived probability that he/she will use the system. Therefore, the rationale behind these relationships is that technology that is easy to use and is found to be particularly useful will have a positive influence on the userâĂŹs intention towards using the technology. The basic TAM model, defined by the three constructs described above, can be extended with additional dimensions which analyze external variables that may affect technology adoption in each specific context.

Further inspiration for the construction of our model was (Diop et al., 2019), which was aimed to study the user acceptance of an Advanced Traveler Information Systems, capable of mitigating traffic congestion and improving road network performance. In this model, three additional variables were included, related to *familiarity* with road network, *quality* of the provided information, and attitude toward diversion (i.e., attitude toward changes in habitual routes, which is often hard for some drivers). The latter was the only additional variable that we did not include in our model, as it was very specific for the application domain.

As far as concerns information quality, this was translated into the quality of the AI results. To measure this, we based our model on the aforementioned work on metrics for explainable AI (Hoffman et al., 2019).

Our proposed **AI Technology Acceptance Model (AI-TAM)** is thus composed of seven basic constructs: perceived usefulness, ease of use, behavioral intention (from TAM), familiarity (which can be measured both towards AI and towards the system-specific application scenario), AI perceived quality and trust in AI (from XAI metrics) and *collaborative intention*. We added the last construct to measure not only the intention towards the use of the AI system but also the willingness to participate âĂIJin-the-loopâĂİ, thus contributing to the continuous improvement of the AI results. Our full AI-TAM model is represented in Figure 1.

All these constructs are the latent dependent variables of the model we analyze in our study; we, however, also aimed to take into account the specificity of AI user experience. Indeed, AI results are in general unpredictable (even when explainable), so a well-performing AI, with respect to a

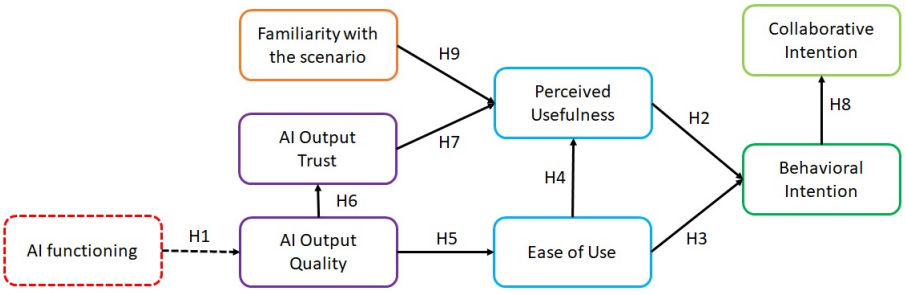

***Figure 1.** Our AI Technology Acceptance Model (AI-TAM). AI functioning is not a construct of the model, it is a control variable used during the model validation.*

poorly performing one, might significantly change AI perceived quality and user trust in AI which, in turn, can have an impact on user acceptance and willingness to act as human-in-the-loop. To this end, in our AI-TAM model, we propose to introduce the *AI correct functioning* as a further control variable, that – whenever suitable and possible – can be employed in an A/B testing design.

To evaluate the AI-TAM model, we propose an assessment questionnaire, mostly composed of 1-to-5 Likert scales, with multiple items for each construct of the model. In particular, the questionnaire includes 5 items for trust in AI results (XAIT), 3 items for perceived quality of AI results (XAIQ), 3 items on perceived usefulness (PUF), 6 items on ease of use (EOU), 6 items for behavioral intention (BI), 3 items for collaborative intention (CI) and 6 items on familiarity (FAM).

The research hypothesis for the validation of the model are the following (reported also in Figure 1):

- H1: The AI functioning (experimental condition) influences the perceived quality in its results (XAIQ)
- H2: The perceived usefulness (PUF) influences the behavioral intention (BI)
- H3: The ease of use (EOU) influences the behavioral intention (BI)
- H4: The ease of use (EOU) influences the perceived usefulness (PUF)
- H5: The quality of the explainable AI results (XAIQ) influences the ease of use (EOU)
- H6: The quality of the explainable AI results (XAIQ) influences the trust in the explainable AI results (XAIT)
- H7: The trust in the explainable AI results (XAIT) influences the perceived usefulness (PUF)
- H8: The intention to use the system (BI) influences the intention to collaborate to its improvement (CI)
- H9: Higher familiarity with the proposed scenario will influence the perceived usefulness of the AI solution (PUF).

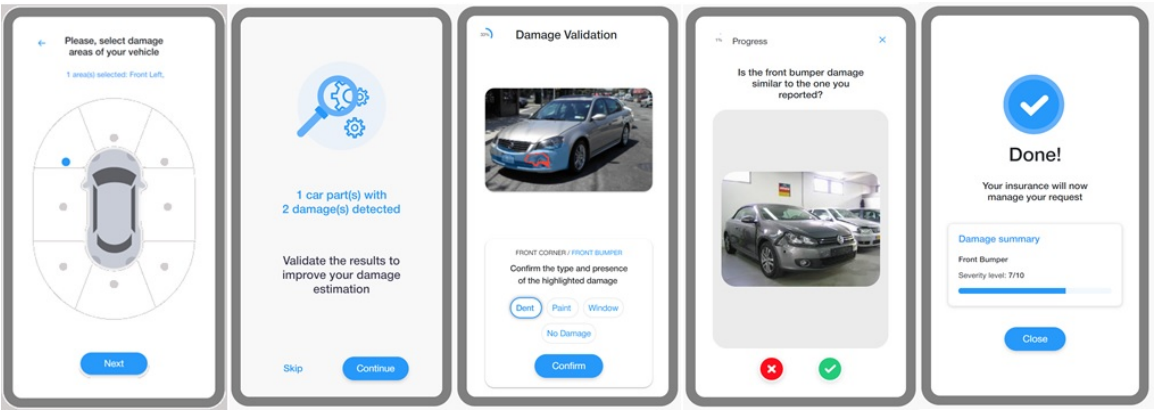

Figure 2. ***Some screenshots of the evaluated application.***

4. USAGE SCENARIO AND EVALUATED SYSTEM

More and more present in everyday applications and services, AI is often employed for the classification of unstructured information like images. In this study, to experiment with the proposed AI-TAM model, we focus on a real application for car accident classification and reporting to insurance named BumpOut (http://bump-out.nl/).

This app provides step-by-step support for an end-user in the correct documentation and claim of a car accident to her insurer, with the AI supporting the analysis of the car damage pictures and in their subsequent severity estimation (see Figure 2). In all the steps of the accident reporting, the user has the possibility to be "in the loop", confirming or correcting the AI identification and classification of the damages (e.g. dent, paint damage, broken window, etc.) and complementing damage information not spotted by the AI. Human contributions are fed back to the AI learning algorithm, to improve it over time and over repeated use.

To evaluate user acceptance and their willingness to be "in the loop", we experimentally evaluated our acceptance model in two steps: a preliminary qualitative/quantitative study, involving 20 subjects in individual "think aloud" sessions, and a second quantitative study, involving 400 subjects through two different crowdsourcing campaigns. The quantitative results of the preliminary study (not reported in this paper), even if not large enough to be statistically valid, were encouraging in that they confirmed our initial hypotheses. In the rest of the paper, therefore, we report only the results of the second and main quantitative study.

The users involved in this study tried an interactive prototype of the application, but it is important to underline that this is not a mere experimental exercise: we used the collected qualitative and quantitative feedback to improve the design of this real-world AI-powered system with a human-in-the-loop approach. Indeed, we designed, developed, and tested the BumpOut application on the basis of the requirements of a European large insurance group, we improved it on the basis of the collected user feedback and now the application is ready for market launch as a standalone app or to be integrated in existing insurance apps.

5. USER EVALUATION METHODOLOGY

For the validation of the model, we set up two crowdsourcing campaigns on the Prolific platform[1] (Palan and Schitter, 2018) involving 400 users during the month of April 2021. For the selection of the crowd workers, the inclusion criteria were: age between 18 and 65, resident in Europe, a driving license and the habit of driving (own or someone elseâĂŹs car), and fluency in English. No inclusion or exclusion criteria were imposed on any other personal characteristic (gender, nationality, etc.). In both campaigns, the users were introduced to an ideal scenario in which, after a car collision, they decided to use the app to understand the damage level and to send the information about the damage to their insurance company. They were invited to test the interactive prototype of the BumpOut application and to report the given car accident from start to finish.

We set up two campaigns (200 users each) in order to employ the control variable of the *AI functioning*. Each campaign has different experimental conditions: the users of the first campaign experienced a flawless AI (i.e., the prototype included a "perfectly functioning" AI, which required

[1] Cf. https://www.prolific.co/

the user only to accept the AI results), and the users of the second campaign experienced a partially failing AI (i.e., the prototype required the user to correct some deliberately added AI mistakes). Section 11.2 of the Supplemental Materials reports the links of the prototypes experienced by the two groups and Section 11.1 shows the introduction and context explained to both groups before experiencing the prototype.

After the interaction with the app prototype, the participants were provided with a questionnaire composed of a set of scales. The first scale corresponded to our AI-TAM-related questions, as illustrated in a previous section; the familiarity was measured both towards technology in general and towards the car accident-specific scenario. The second scale in the questionnaire was the System Usability Scale (SUS) (Brooke, 1996): a 10 item scale used to evaluate system usability. We added also an item to compute the collective Net Promoter Score (NPS) (Reichheld, 2004) with its usual 1-to-10 Likert scale. The complete list of items presented to the participants is provided in the Supplemental Materials in section 11.3. Finally, it is important to underline that no incentive mechanisms were used. The first reason is that the recruitment on crowdsourcing campaigns already includes economic compensation. Furthermore, the study wants to collect information without adding any bias in the naive behavioral intention of the final users. The inclusion of further incentive mechanisms could be analyzed in future studies.

6. PRELIMINARY ANALYSIS

We first assessed the reliability of the questionnaire used to collected the data using Cronbach's alpha (Brown, 2002). We obtained an alpha value above the cut-off value ($\alpha > 0.7$) for all the constructs on which the index was applicable (construct with at least 3 questions and with the same scale). This means that there is a high internal consistency and the items of each construct are intercorrelated and produce consistent responses. With respect to the two experimental conditions, in the following we name "FlawlessAI-Group" the subset of users who experienced a flawless AI and "FailingAI-Group" the one with partially failing AI.

The two groups are homogeneous in terms of familiarity with the technology and with the accident scenario, since the differences between the answers collected in the two groups are not statistically significant at 5% significance level. The two groups are also balanced in terms of the participants' age with an average of 25 years for both groups.

Regarding the *familiarity with technologies*, the participants of both groups are very confident in using the smartphone or the apps in general (M=4.79, SD=0.55 for the FlawlessAI-Group and M=4.67, SD=0.54 for the FailingAI-Group on a 1-5 scale, independent t-test t(397)=2.19, p=0.029), but they are a bit less confident in using AI applications (M=3.47, SD=0.95 for the FlawlessAI-Group and M=3.37, SD=0.84 for the FailingAI-Group on a 1-5 scale, independent t-test t(392)=1.055, p=0.292). 25% of them reported that they do not daily use AI-powered systems and 50% declared to use 1 or 2 AI-powered apps every day.

Regarding the *familiarity with the car accident scenario*, the involved users were not very experienced: 60% of respondents never had an accident while driving (M=2.19, SD=0.53 for the FlawlessAI-Group and M=2.16, SD=0.53 for the FailingAI-Group on a 1-5 scale, independent t-test t(397)=0.5647, p=0.572). Half of them never supported a friend or relative after they had an accident (both in case of presence during the accident and of a later arrival after the accident

occurred) and the other half gave support 1 or 2 times. In case of a car accident, only half of the participants declared to be confident in reporting it to the insurance company by themselves (M=3.52, SD=1.16 for the FlawlessAI-Group and M=3.38, SD=1.19 for the FailingAI-Group on a 1-5 scale, independent t-test t(397)=1.22, p=0.22).

To verify hypothesis H9, we computed the Pearson's correlation between the familiarity and perceived usefulness variables; since the two groups were very similar with respect to the familiarity dimension, we evaluated it on the entire set of participants. The Pearson's correlation coefficient is very low and not statistically significant, both between familiarity with the technology and perceived usefulness ($r(398)=0.1$, p=0.038), and between familiarity with accidents and perceived usefulness ($r(398)=0.04$, p=0.417). For this reason, we reject H9, in that we do not have sufficient evidence that the familiarity with the application scenario/characteristics influence the perceived usefulness.

To complete the analysis on familiarity, we also tested its correlation with the other AI-TAM model constructs, but we did not find any medium or large association (r ranging between 0 and 0.2) and none of the correlations was statistically significant. This seems to suggest that the familiarity variable does not affect any other construct for the BumpOut application. A possible interpretation is that this app is very intuitive and thus its value is perceived independently from previous experiences with car accidents or technology. For those reasons, in the rest of the paper, we will not consider again the familiarity construct in the further evaluations of our model.

In order to test hypothesis H1, we evaluated the interplay between the control variable (AI functioning) and the perceived quality of AI results (XAIQ). A t-test for the difference in mean of the XAIQ items between the two groups yield a significant result, but the two means are actually very close (M=4.07, SD=0.53 for the FlawlessAI-Group and M=3.86, SD=0.57 for the FailingAI-Group on a 1-5 scale, independent t-test t(396)=3.77, p=0.0001), as it can be appraised from the box-plot in Figure 3.

	Gr. 1 Mean	**Gr. 1 Std dev**	**Gr. 2 Mean**	**Gr. 2 Std dev**	**t-test**
XAIQ	4.07	0.53	3.86	0.57	t(396)=3.7726, p=0.0001 ***
XAIT	3.84	0.63	3.76	0.65	t(397)=1.2314, p=0.21
PUF	4.26	0.64	4.05	0.71	t(393)=3.0384, p=0.002 **
EOU	4.61	0.47	4.44	0.62	t(370)=3.2043, p=0.001 ***
BI	3.83	0.71	3.70	0.72	t(397)=1.8106, p=0.07
CI	4.25	0.65	4.09	0.71	t(395)=2.3875, p=0.01 **

***Table 1.** Mean and standard deviation for each construct and t-test for difference in mean between the two experimental conditions (asterisks indicate significance level: *** p-value<0.001, ** p-value<0.01, * p-value<0.05).*

We extended this analysis on the effects of the AI functioning to the other constructs: trust in AI results (XAIT), perceived usefulness (PUF), perceived ease of use (EOU), behavioral intention (BI) and collaborative Intention (CI). Table 1 shows for each construct the mean and standard deviation of each group and the p-value of the t-test for difference in mean (all item values range from 1 to 5). The differences are statistically significant at least at a 1% significance level for all the

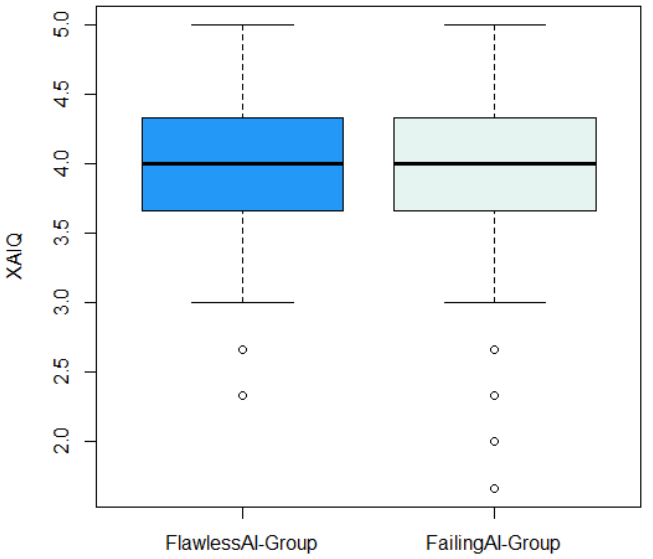

***Figure 3.** Boxplot of XAIQ scores of the two groups*

constructs except for BI and XAIT. However, those differences in mean between the two groups are in general very limited. In other words, we can accept the hypothesis âĂIJH1: The AI functioning (experimental condition) influences the perceived quality in AI results (XAIQ)âĂİ, as a small effect is noticeable, therefore the two groups will be merged in a single one in the following analysis. At conceptual level, this result could be justified by the usersâĂŹ expectations: if they are willing to cooperate with the app (and so with the AI) to achieve the result, the need to correct the output does not make a big difference from their perception.

As a preliminary evaluation of the rest of the hypotheses, we computed the Pearson's correlations between all constructs, which are shown in Table 2.

XAIT and XAIQ constructs appear to be strongly correlated one with the other ($r(398)=0.65$, $p<0.001$) and also with PUF, BI and EOU (r ranging between 0.51 and 0.69). PUF is also strongly correlated with EOU ($r(398)=0.60$, $p<0.001$) and BI ($r(398)=0.59$, $p<0.001$), in accordance with the TAM model. The lowest correlation values, still showing a medium strenght of the associations, are recorded between the CI construct and the other variables, with the highest value between CI and BI ($r(398)=0.45$, $p<0.001$). These preliminary results show that indeed our AI-TAM model – with the exception of the familiarity construct, as explained before – seem to be valid for the evaluated application. Therefore, in the next section, we deepen our investigation with a confirmatory factor analysis.

	XAIT	XAIQ	PUF	EOU	BI
XAIQ	0.65				
PUF	0.69	0.60			
EOU	0.51	0.51	0.60		
BI	0.67	0.54	0.59	0.43	
CI	0.41	0.34	0.44	0.44	0.45

Table 2. ***Pearson's correlations between constructs (r values). All coefficients are statistically significant at a 0.1% significance level.***

Before proceeding with the rest of the study, we can still provide some other global considerations. The users of both groups gave high scores to the items related to the perceived usefulness and the ease of use constructs (cf. Table 1). This indicates that the BumpOut application is perceived as a useful tool that could help users in a easy way, quickly and effectively. The interaction with the app is clear and understandable and it is easy to learn how to use it. The high scores in the XAIT and XAIQ constructs also indicate that users trust the app and they are confident that it works well and its output are reliable. Both user groups declared the intention of using the app in a real situation of a car accident (BI means of 3.83 and 3.70 respectively) as well as the willingness to cooperate to improve the damage detection mechanism within the human-in-the-loop mechanism (CI means of 4.25 and 4.09 respectively).

We assessed the usability of the tool also with the already mentioned System Usability Scale (SUS), whose score computed on the entire dataset is good (70.6, over the 70 arbitrary value usually adopted as threshold), with a slightly higher value for the FlawlessAI-Group (71.6 vs 69.7, independent t-test t(381)=2.2785, p=0.023).

Finally, with respect to the Net Promoter Score (NPS), there is no significant difference in the two experimental conditions. The final total NPS is 8, which indicates a slightly positive attitude towards "promotion".

7. CONFIRMATORY FACTOR ANALYSIS RESULTS

As explained in the previous section, given the limited differences between the two experimental conditions, we aggregated the collected data from the two crowdsourcing campaigns in a single dataset (cf. Supplemental Materials Section 11.4), which was analyzed with a Confirmatory Factor Analysis (CFA) (Brown, 2015), by using R with the Lavaan package version 0.6-8. We evaluated the goodness of fit for our AI-TAM model with the most commonly used fit indexes (Kline, 2015): the comparative fit index (CFI), the Tucker Lewis Index (TLI), the standardized root mean square residual (SRMR) and the root mean square error of approximation (RMSEA). The usual thresholds for those measures are 0.9 (or above) for CFI and TFI and 0.05 (or below) for SRMR and RMSEA.

We built the initial model by considering every construct of the AI-TAM (except for familiarity) as a latent variable, described by the items of the respective scales (considered as observed indicators or manifest variables). Furthermore, regression was used among latent variables, directed as described in the hypotheses. This model is shown in Figure 4, with the resulting CFA loadings on the arrows.

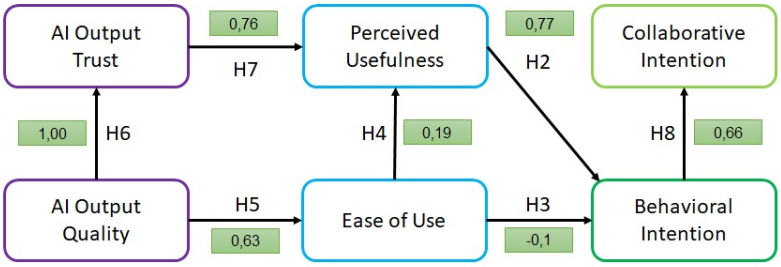

Figure 4. ***Results of the CFA of the initial model, considering regression between latent factors. Numbers on the arrows indicate the regression loadings.***

Overall, the measures for the goodness of fit of this model were not extremely good (CFI=0.906, TLI=0.895, RMSEA=0.063, and SRMR=0.076), since the TLI is below the recommended threshold of 0.9. Moreover, the loadings between EOU and PUF and between EOU and BI were respectively very low (0.19) and negative (-0.1), below the 0.3 usually adopted threshold. On the other hand, the loading between XAIQ and XAIT was exactly 1, indicating that the two constructs cannot be completely distinguished. Indeed, an exploratory factor analysis (EFA) on the entire dataset identified XAIT and XAIQ as a unique construct, therefore, in the continuation of our analysis, we merged the respective items in a single XAI construct.

In order to attain a better fit, we re-specified the model exploiting the Modification Indices (MIs). MIs give an indication of the parameters to be removed or added to the model in order to improve the goodness of fit measures. This is a data-driven modification of the original hypothesized model that can lead to a better fitting model.

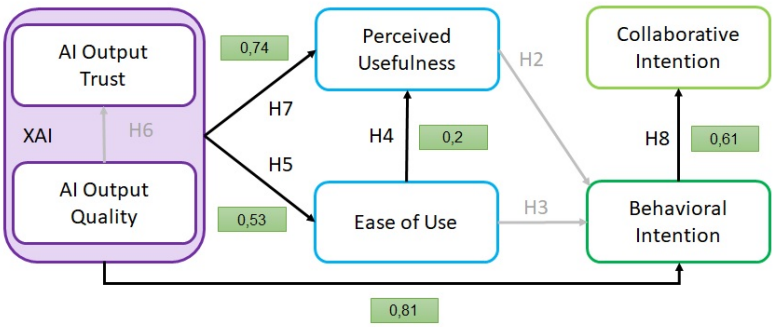

Figure 5. ***Final CFA model, with the indication of the accepted and rejected hypotheses as well as the regression loading values.***

With this technique, we obtained the model shown in Figure 5, which had a substantial and decisive improvement on the goodness of fit measures (CFI=0.974, TLI=0.970, RMSEA=0.035, and RMSR=0.049), with the CFI and TLI well above the recommended threshold of 0.9 and RMSEA

and RMSR correctly below the threshold of 0.1. Both initial and final models, with their detailed graphical representations, are also available in the Supplemental Materials in Section 11.4.

In this final model, the XAIQ5 item ("The app is efficient in that it seems to work very quickly.") is also a manifest variable of the EOU construct, while items EOU2 ("It seems easy to get this app to do what I want it to do.") and EOU4 ("I find this app to be flexible enough to interact with it") are manifest variables also of the XAI construct. From those items' formulations, this modification seems indeed reasonable.

The regressions between latent variables identified by the model are as follows: the XAI construct predicts behavioral intention (BI, loading 0.81), perceived usefulness (PUF, loading 0.74) and ease of use (EOU, loading 0.53); collaborative intention (CI, loading 0.61) is predicted by behavioral intention; only the influence of EOU on PUF is limited (loading 0.20).

8. DISCUSSION

Considering the final model shown in Figure 5, confirmed by the fit measures, we can proceed to accept or reject our AI-TAM model hypotheses for the BumpOut application.

Hypotheses H2 (The PUF influences the BI) and H3 (The EOU influences the BI) are rejected by the final model since the regressions between PUF and BI and between EOU and BI were eliminated. For H2, relevant evidence was found while evaluating the first model with a significant regression coefficient (0.77), while for H3, also in the original model, despite a medium Pearson's correlation coefficient (0.43), the regression coefficient was negative and close to zero (-0.1). In the final model, it seems that BI is mostly influenced by the XAI latent variable (0.81), which maybe "shadows" the direct effect of PUF and EOU.

Hypothesis H4 (The EOU influences the PUF) can be at least partially accepted, as both in the initial and final model there is a significant effect even if the regression coefficient between EOU and PUF is not very strong (0.20 in the final model). The interpretation here is that ease of use only partially contributes to the perception of usefulness.

Hypotheses H7 (The XAIT influences the PUF) and H5 (The XAIQ influences the EOU) can be accepted as well, as can be seen in the final model by looking at the high regression coefficients of the XAI latent variable on PUF and EOU (0.74 and 0.53 respectively). The only difference from the original theory is that XAIT and XAIQ were merged in the final model, as they provided information about the same latent construct in the case of the BumpOut application.

For the same reason, no conclusions can be drawn from the final model for hypothesis H6 (The XAIQ influences the XAIT): in the original model, the regression coefficient between XAIT and XAIQ was 1, which contributed to the decision to merge the two latent variables into a single construct. Further evidence is needed to understand if the AI perceived quality and the trust in AI are indeed two different factors in the evaluation of this kind of application.

Hypothesis H8, investigating the dependency of CI from BI, found strong evidence both in the original model and in the final model, with regression coefficients of similar weights (0.66 and 0.61 respectively). Therefore, in the case of BumpOut, we can indeed conclude that if a user is willing to adopt the app she/he will also likely be happy to be involved in the human-in-the-loop feedback activities. This is a very nice result, as feedback loops usually put a heavier burden in terms of

tasks required to the user: we can conclude that the design of the BumpOut app took into due consideration the mental model and expectations of its intended users, resulting in a high behavioral intention, which in turn contributed to a consistent collaborative intention.

The main difference between the original model and the final model is that the AI-related constructs (independently from the fact that AI might sometimes fail) seem to have fundamental importance for all the original TAM constructs, i.e. perceived ease of use (EOU), perceived usefulness (PUF), as well as behavioral intention to use this system (BI). Indeed, an additional and quite strong influence was found between XAI and BI (regression coefficient of 0.81). This suggests that the introduction of AI can have a non-negligible effect on the user propensity to adopt an application and, therefore, this is a phenomenon worth investigating and evaluating.

9. CONCLUSIONS AND NEXT STEPS

In this paper, we proposed a model to evaluate the user acceptance of AI-powered applications with a human-in-the-loop mechanism and we experimentally tested it on an application for the reporting and automatic damage estimation of car accidents.

This research is only a first evaluation of our proposed AI-TAM model, in the context of a specific user application. However, we believe that this kind of investigation can become more and more crucial nowadays, with the ever-increasing introduction of AI features in everyday systems.

Moreover, with the need to involve the final user in AI pipelines according to a human-in-the-loop paradigm, user acceptance becomes even more fundamental. This is the reason why we introduced the collaborative intention construct: a user could be willing to use a specific system or technology (behavioral intention), but she/he may not necessarily have the same propensity to "cooperate" with the machine to give it feedback or contributions to improve.

The presented study showed that the AI-related constructs of our AI-TAM model have a strong and positive effect on the behavioral intention, the perceived usefulness, and the ease of use of the application (the main latent factors of the original TAM). Moreover, there is a strong link between behavioral intention and collaborative intention, indicating that indeed human-in-the-loop approaches can be successfully adopted in final user applications.

The results of the presented evaluation were instrumental to the design improvement of the BumpOut application, which was adapted to take into account the results of the performed qualitative assessment. On the other hand, we were pleased by the consistently high values of AI perceived quality and trust in AI and the limited impact of the AI functioning control condition in the two experimental conditions. This let us continue the app development with the confidence that indeed such human-in-the-loop mechanism can be introduced in a car damage claiming app, with the positive collateral effect to enable an insurance company to collect additional user-provided information to improve the adopted AI algorithm.

Our future work will be devoted to further applying the proposed AI-TAM model to evaluate other AI-powered systems, especially when adopting a human-in-the-loop paradigm, in order to refine and possibly extend the model.

ACKNOWLEDGMENTS

The authors would like to thank all the participants that took part in the crowdsourcing campaign. This research was partially funded by EIT Digital in the AIDE project (Activity Code: 19386).

10. REFERENCES

AI HLEG, . (2019). Ethics Guidelines for Trustworthy AI. (2019).

AI HLEG, . (2020). The Assessment List for Trustworthy Artificial Intelligence (ALTAI). (2020). DOI:http://dx.doi.org/10.2759/002360

Amershi, S, Fogarty, J, and Weld, D. (2012). Interactive Machine Learning for On-Demand Group Creation in Social NetworksâĂŹ. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 21–30.

Arya, V, Bellamy, R. K, Chen, P.-Y, Dhurandhar, A, Hind, M, Hoffman, S. C, Houde, S, Liao, Q. V, Luss, R, Mojsilović, A, and others, . (2019). One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* (2019).

Baker-Brunnbauer, J. (2009). *Trustworthy AI Implementation (TAII) Framework for AI Systems*. Technical Report. SSRN. DOI: http://dx.doi.org/10.2139/ssrn.3796799

Brooke, J. (1996). Sus: a âĂIJquick and dirtyâĂŹusability. *Usability evaluation in industry* 189 (1996).

Brown, J. D. (2002). The Cronbach alpha reliability estimate. *JALT Testing & Evaluation SIG Newsletter* 6, 1 (2002).

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.

Cahour, B and Forzy, J.-F. (2009). Does projection into use improve trust and exploration? An example with a cruise control system. *Safety Science* 47, 9 (2009), 1260–1270. DOI:http://dx.doi.org/https://doi.org/10.1016/j.ssci.2009.03.015 Research in Ergonomic Psychology in the Transportation Field in France.

Davis, F. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly* (1989), 319–340.

Dietvorst, B. J, Simmons, J. P, and Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology. General* 144 1 (2015), 114–26.

Diop, E. B, Zhao, S, and Duy, T. V. (2019). An extension of the technology acceptance model for understanding travelersâĂŹ adoption of variable message signs. *PLOS ONE* 14, 4 (04 2019), 1–17. DOI:http://dx.doi.org/10.1371/journal.pone.0216007

Dodge, J, Liao, Q. V, Zhang, Y, Bellamy, R. K. E, and Dugan, C. (2019). Explaining models: an empirical study of how explanations impact fairness judgment. *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019).

European Commission, . (2021). Artificial Intelligence Act (Proposal for a Regulation of the European Parliament and of the Council "Laying down harminized rules on Artificial Intelligence" - COM(2021). (2021).

Fishbein, M and Ajzen, I. (1977). Belief, attitude, intention, and behavior: An introduction to theory and research. *Philosophy and Rhetoric* 10, 2 (1977).

Fryer, D. (1939). Post Quantification of Introspective Data. *The American Journal of Psychology* 52, 3 (1939), 367–371. http://www.jstor.org/stable/1416744

Green, B and Chen, Y. (2019). The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 50 (Nov. 2019), 24 pages. DOI:http://dx.doi.org/10.1145/3359152

Hallensleben, S and others, . (2009). *From Principles to Practice - An interdisciplinary framework to operationalise AI ethics*. Technical Report. AI Ethics Impact Group.

Hidalgo, C. A, Orghiain, D, Canals, J. A, De Almeida, F, and Martin, N. (2021). *How humans judge machines*. MIT Press.

Hoffman, R. R, Mueller, S. T, Klein, G, and Litman, J. (2019). Metrics for Explainable AI: Challenges and Prospects. (2019).

Honeycutt, D, Nourani, M, and Ragan, E. (2020). Soliciting Human-in-the-Loop User Feedback for Interactive Machine Learning Reduces User Trust and Impressions of Model Accuracy. (08 2020).

Howe, J. (2008). *Crowdsourcing: How the power of the crowd is driving the future of business*. Random House.

Jian, J.-Y, Bisantz, A. M, and Drury, C. G. (2000). Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

K.E., S. (2013). The perception and measurement of human-robot trust. *Doctoral dissertation, University of Central Florida Orlando, Florida* (2013).

Klein, G and Hoffman, R. (2008). Macrocognition, mental models, and cognitive task analysis methodology. *Naturalistic Decision Making and Macrocognition* (01 2008), 57–80.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Law, E and Ahn, L. v. (2011). *Human computation*. Vol. 5. Morgan & Claypool Publishers. 1–121 pages.

Longoni, C, Bonezzi, A, and Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research* 46, 4 (2019), 629–650.

Luger, E and Sellen, A. (2016). " Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5286–5297.

Madsen, M and Gregor, S. (2000). Measuring human-computer trust. In *Proceedings of the 11 th Australasian Conference on Information Systems*. 6–8.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

Mittelstadt, B, Russell, C, and Wachter, S. (2019). Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*. 279–288.

Palan, S and Schitter, C. (2018). Prolific.ac – A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27.

Ray, A, Burachas, G, Yao, Y, and Divakaran, A. (2019). Lucid Explanations Help: Using a Human-AI Image-Guessing Game to Evaluate Machine Explanation Helpfulness. (04 2019).

Reichheld, F. (2004). The One Number you Need to Grow. *Harvard business review* 81 (06 2004), 46–54, 124.

Ribeiro, M. T, Singh, S, and Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

Settles, B. (2009). *Active learning literature survey*. Technical Report. University of Wisconsin-Madison Department of Computer Sciences.

Thiebes, S, Lins, S, and Sunyaev, A. (2020). Trustworthy artificial intelligence. *Electronic Markets* (2020). DOI:http://dx.doi.org/10.1007/s12525-020-00441-4

Wang, R, Harper, F. M, and Zhu, H. (2020). Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020).

Williams, M, Hollan, J, and Stevens, A. (1983). Human Reasoning About a Simple Physical System. In *Mental models*.

Zicari, R. V, Brodersen, J, Brusseau, J, DÃijdder, B, Eichhorn, T, Ivanov, T, Kararigas, G, Kringen, P, McCullough, M, MÃűslein, F, Mushtaq, N, Roig, G, StÃijrtz, N, Tolle, K, Tithi, J. J, van Halem, I, and Westerlund, M. (2021). Z-InspectionÂő: A Process to Assess Trustworthy AI. *IEEE Transactions on Technology and Society* 2, 2 (2021), 83–97. DOI:http://dx.doi.org/10.1109/TTS.2021.3066209

11. SUPPLEMENTAL MATERIALS

The supplemental materials include the introductory text given to the users in the crowdsourcing campaign to explain them how the app works and the scenario they should imagine while trying the prototype (Section 11.1). Section 11.2 lists the links of the two prototypes and Section 11.3 shows the list of questions included in the questionnaire with the corresponding construct. The answers collected and the model fitted with this data are reported in Section 11.4. All the files referenced in this paper are packaged as a Research Object, openly available for reproducibility on Zenodo, permanently identified by the DOI 10.5281/zenodo.6541969.

11.1. The context

Why this app? The app was designed to help people who were involved in an accident with their car. The goal is to support users during these stressful, vulnerable and uncertain moments by providing a very simple interface to guide the user step by step through the damage claiming process.

The app helps users submitting the pictures of the car damages, so that the claiming process can start in a very simple way and directly from their digital devices. This can help users in speeding up the claiming process without wasting time in arranging appointments with the insurance experts.

How does it work? The app flow is composed of these 5 main steps:

i. The user starts a new damage claim and inserts the car information
ii. The user selects all the damaged car regions and, for each region selected, he/she uploads a picture of the damaged areas
iii. The app analyzes the images and identifies both the damaged car regions and the damage areas itself. The user can give his/her feedback by confirming the presence and the type of the damages or add missing damages not identified by the app.
iv. To better compute the final estimation, the app asks the user to select the images that displays a damage similar to the one they are claiming
v. At last, the list of broken car regions with an estimation of the severity level is displayed to the user.

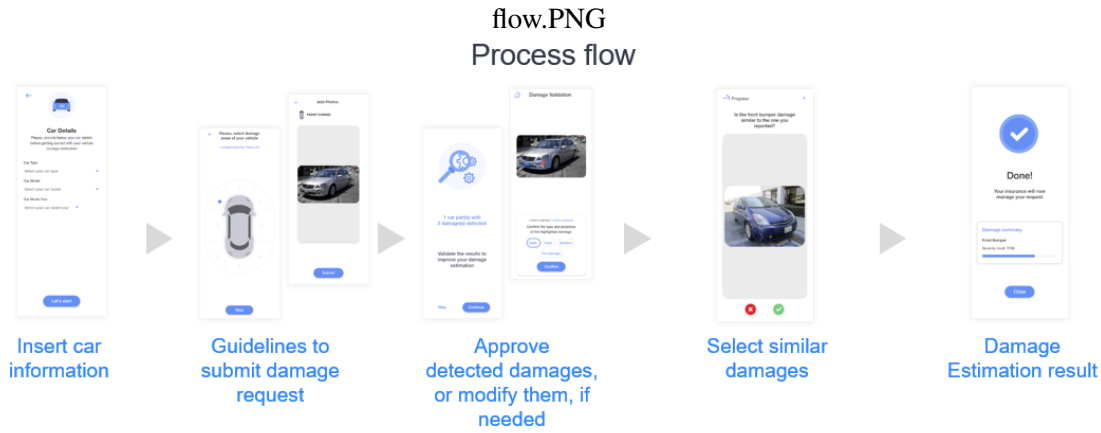

Figure 6. *The 5 main steps of the process*

Scenario So, now let's imagine to have a car collision, and some parts of your car are damaged. To understand the damage level and to send the right information about the damage to your insurance company, you decided to try this digital platform.

11.2. The interactive prototype

The links to the interactive prototypes experienced by the two groups:

– FlawlessAI-Group: https://bit.ly/bo-prototype-flawlessAI
– FailingAI-Group: https://bit.ly/bo-prototype-failingAI

11.3. The questions

Full list of items of the user questionnaire. For each question the corresponding construct and the available answers are reported (file "list-questions.pdf" in the Research Object).

Here is a list of the constructs abbreviations:

– FAM-ACC: Familiarity with the Accident scenario
– FAM-TEC: Familiarity with Technology
– XAIT: Trust in AI results
– XAIQ: Perceived Quality of AI results
– PUF: Perceived Usefulness
– EOU: Ease Of Use
– SUS: System Usability Scale
– BI: Behavioral Intention
– CI: Collaborative Intention
– NPS: Net Promoter Score

Construct	Question	Answer values
FAM-ACC-1	Have you ever had an accident while you were driving?	No - from 1 to 2 times âĂŞ From 2 to 4 times âĂŞ More than 4 times
FAM-ACC-2	Have you ever supported a friend or relative after they had an accident (both in case of presence during the accident, and if you arrived after the accident was already occurred, as a support)?	No - from 1 to 2 times âĂŞ From 2 to 4 times âĂŞ More than 4 times
FAM-ACC-3	In case of a car accident, how confident do you feel to report it to the insurance company by yourself?	1-to-5 Likert scale
FAM-TEC-1	How much do you feel confident in using the smartphone or apps?	1-to-5 Likert scale
FAM-TEC-2	How confident do you feel with Artificial Intelligence applications?	1-to-5 Likert scale
FAM-TEC-3	How many times per day do you use applications that use Artificial Intelligence?	I do not use them - From 1 to 2 times - From 2 to 4 times - More than 4 times
XAIT-1	I would be confident in the app. I feel that it works well	1-to-5 Likert scale
XAIT-4	I feel that, by relying on the app, I will get the right answers.	1-to-5 Likert scale
XAIT-6	I tend not to trust the app.	1-to-5 Likert scale
XAIT-7	It seems that the app can perform the task better than a novice human user.	1-to-5 Likert scale
XAIT-8	I would like to use the app for decision making.	1-to-5 Likert scale
XAIQ-2	The outputs of the app are very predictable	1-to-5 Likert scale
XAIQ-3	The app is very reliable. I could count on it to be correct all the time.	1-to-5 Likert scale
XAIQ-5	The app is efficient in that it seems to work very quickly.	1-to-5 Likert scale
PUF-1	Using this app would allow me to accomplish the related tasks more quickly.	1-to-5 Likert scale
PUF-2	Using this app would enhance my effectiveness on the tasks related with its usage.	1-to-5 Likert scale
PUF-3	Using this app would make it easier to do actions connected to its usage.	1-to-5 Likert scale
EOU-1	Learning to operate this app would be easy for me.	1-to-5 Likert scale
EOU-2	It seems easy to get this app to do what I want it to do.	1-to-5 Likert scale

EOU-3	The interaction with this app is clear and understandable.	1-to-5 Likert scale
EOU-4	I find this app to be flexible enough to interact with it.	1-to-5 Likert scale
EOU-5	It would be easy for me to become skilled at using this app.	1-to-5 Likert scale
EOU-6	This app seems easy to be used	1-to-5 Likert scale
SUS-1	I think that it would help me to use this app when an accident happens.	1-to-5 Likert scale
SUS-2	I found this app unnecessarily complex.	1-to-5 Likert scale
SUS-3	I think that I would need the support of a technical person to be able to use this app.	1-to-5 Likert scale
SUS-4	I found the various functions in this app were well integrated.	1-to-5 Likert scale
SUS-5	I thought there was too much inconsistency among the app functionalities.	1-to-5 Likert scale
SUS-6	I imagine that most people would learn to use this app very quickly.	1-to-5 Likert scale
SUS-7	I found the app very cumbersome to use.	1-to-5 Likert scale
SUS-8	I think I would feel very confident using this app.	1-to-5 Likert scale
SUS-9	I needed to learn a lot of things before I could get going with this app.	1-to-5 Likert scale
BI-1	I think I would like to install this app in advance, to be prepared in case of car accident.	1-to-5 Likert scale
BI-2	In case of an accident, I would document it in a traditional way, without the app	1-to-5 Likert scale
BI-3	Even if not installed in my phone, in case of an accident I think I would download and use this app.	1-to-5 Likert scale
BI-4	I think that in case of a car accident I would use this app.	1-to-5 Likert scale
BI-5	In an accident with another car, if the other driver wants to claim the damage with this app, I would agree to use it.	1-to-5 Likert scale

BI-7	What features would you like to be added to the app before putting it into operation? Select max 3 options	Repair shops affiliated with the insurance - Economic evaluation range for estimating the damage - Information on insurance coverage - Information about next steps - Free text area to add comments about claim - Reminder of legal/administrative deadlines - Repair times - My case number - Information about the meaning of the severity level (5/10) - None
NPS	On a scale of 1 to 10, how likely are you to recommend this app to a friend or colleague?	1 - 10 scale
CI-1	In your opinion, does the damage detection mechanism need improvements?	1-to-5 Likert scale
CI-2	Would you find it useful if people would give their contribution to help the app in improve damage detection?	1-to-5 Likert scale
CI-3	Would you use this app to help the damage detection mechanism to improve?	1-to-5 Likert scale

11.4. **The answers and the model**

The interested reader can find further information about the answers collected and the model within the already mentioned Research Object with DOI 10.5281/zenodo.6541969, as follows:

- Full dataset containing the collected answers: "dataset.csv" file
- Final CFA model specification in Lavaan syntax: "model-syntax.pdf" file
- Detailed graphical representations of the initial and final models (including the loadings on manifest variables): "models-plot.pdf" file