Trust and Perceived Control in Burnout Support Chatbots

Degachi, Chadha; Al Owayyed, Mohammed; Tielman, Myrthe Lotte

DOI
10.1145/3544549.3585780

Publication date
2023

Document Version
Final published version

Published in
CHI 2023 - Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems

Trust and Perceived Control in Burnout Support Chatbots

Chadha Degachi
cdegachi@sigchi.org
Delft University of Technology
Delft, Netherlands

Mohammed Al Owayyed
M.AlOwayyed@tudelft.nl
Delft University of Technology
Delft, Netherlands

Dr. Myrthe Tielman
m.l.tielman@tudelft.nl
Delft University of Technology
Delft, Netherlands

ABSTRACT

Increased levels of user control in learning systems is commonly cited as good AI development practice. However, the evidence as to the effect of perceived control over trust in these systems is mixed. This study investigated the relationship between different trust dimensions and perceived control in postgraduate student burnout support chatbots, and modelled the moderating factors therein. We present an in-between subject controlled experiment using simulated therapy-goal learning to study the effect of perceived control (as manipulated by feedback incorporation) on perceived agent benevolence, competence, and trust. Our results showed that perceived control was moderately correlated with benevolence ($r = 0.448$, $BF_{10} = 7.150$), and weakly correlated with competence and trust.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**.

KEYWORDS

trust modelling, percieved control, human-in-the-loop, chatbots

ACM Reference Format:
Chadha Degachi, Mohammed Al Owayyed, and Dr. Myrthe Tielman. 2023. Trust and Perceived Control in Burnout Support Chatbots. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3544549.3585780

1 INTRODUCTION

eHealth solutions have appeared in recent years hoping to bridge the gap between individuals and mental health services, reaching a population not currently served by in-person support. Unlike traditional approaches, eHealth solutions offer greater scalability, lower cost, anonymity, and resource equity [6, 62]. With rising trends in mental health issues among university students [35, 39, 46] such systems have become vital, especially given the outbreak of the COVID-19 virus and subsequent pandemic [4, 56]. Among these rising mental health issues is burnout [3, 47, 58], a psychological syndrome in which an individual suffers emotional exhaustion, depersonalization, and reduced personal accomplishment [37]. eHealth systems do have their drawbacks however, they tend to suffer from attrition, i.e., the loss of user engagement over time [7, 12, 45]. One method of increasing engagement is the use of chatbots [45]. Chatbots have a long history in mental health care, where they can mimic the support of healthcare professionals, thereby fostering a stronger sense of accountability in users, and promoting engagement [41, 45].

We investigated trust in post-graduate student burnout support chatbots. Trust is a key aspect not only of human-chatbot interactions, but also of human-therapist interactions [34, 49]. In the field of human-computer interaction (HCI), its most common dimensions are benevolence (the confidence that one's wellbeing will be protected by the trustee [19, 26]) and competence (confidence in the trustee's skill level [19, 26]). Improvement along those dimensions creates more positive interactions with technology [38], while maintaining steady relationships with intelligent systems [63].

Many factors influence trust, thus, though we focused on perceived control as manipulated by feedback incorporation in this study, we also modelled possible moderating factors in the trust-control relationship. The evidence as to the exact effect of control over an interaction with an intelligent system is mixed. In some cases, allowing users to correct mistakes made by the system was seen to improve user trust [20, 54, 55], but in others, the opposite is true [24]. Nonetheless, allowing the user some degree of control over intelligent systems remains a recommended 'best-practice' in industry standards [1, 17, 18]. Such discord further motivates the development of a more nuanced understanding of perceived control. Given this, we propose the following research questions:

(1) **RQ1**: How does allowing the user to feedback the agent's predictions affect their perception of its benevolence?
(2) **RQ2**: How does allowing the user to feedback the agent's predictions affect their perception of its competence?
(3) **RQ3**: What factors moderate the relationship between different trust dimensions and perceived control?

To answer these questions, we designed a prototype of a therapy-goal-generating empathetic chatbot. While the chatbot was presented to participants as able to learn goals from conversation, goal formulation was simulated using an abbreviated clinical burnout inventory. We chose goals as the locus of exercising control as goal-setting and goal-alignment are vital activities when establishing trust in client-therapist relationships [25, 43], as well as when motivating and sustaining behavioural change in users [36]. Users exercised control over the chatbot by accepting, rejecting, or correcting the goals it recommended.

2 RELATED WORK

Lee and See [32] define trust in automation as *"[the belief] that an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability"*. Along with benevolence and competence [26], further dimensions such as reliability and utility [23] emerge in human-computer trust. As for perceived control, we

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3585780

define it as the perception of directly altering system behaviour, models, or outcomes. This definition derives from human-in-the-loop machine learning (IML). In IML, users influence model behaviour, most commonly by labelling data points presented to, or selected by, them [59]. The effect of this feedback incorporation technique on perceived control, and by extension trust, is not extensively studied, and studies that address this issue contradict in their findings.

Gutzwiller and Reeder [20] and Smith et al. [54] investigated this relationship in unmanned vehicles and topic modelling respectively, concluding that control via feedback incorporation was positively correlated with trust. In the first study, researchers found that participants not only trusted and preferred the IML regardless of interaction level, but were also able to recognize them as such when compared to other systems. In [54], researchers noted that participants seemed to even overtrust the system. On the other hand, using a simulated face detection model, Honeycutt et al. [24] found correcting the system to be negatively correlated with trust and perception of accuracy, regardless of whether the model improved in accuracy after feedback. Nonetheless, the authors still pointed out that users have shown a higher willingness to use imperfect systems when they were able to correct them [11]. We further note that users may also value systems more if they participated in their training [42]. Moreover, in human teams, feedback provision and incorporation improves the perceived fairness of team decisions and trust in the decision-making process [29, 61]. Of these studies, both Honeycutt et al. [24] and Smith et al. [54] choose to select non-expert end-users as their participant pool, contrasting their choice with the more common choice of developers or annotators. Similarly, we were also motivated to involve the real-world use case of student burnout, especially as it allowed us to incorporate the elements of vulnerability into the system-user interaction that are vital to trust formation.

Comparison across these studies is somewhat hindered as they do not employ the same scale for measuring trust, nor do they employ granular measures of trust as we have. However, Honeycutt et al. [24] do invoke themes of competence in their discussion, speculating whether the negative impact on trust was due to the action of system correcting embedding system mistakes more deeply in the users' memory than its successes. Thus, within the context of the postgraduate student population, we propose the following hypotheses:

(1) **H1**: Increased perceived control has a positive correlation with perceived benevolence.
(2) **H2**: Increased perceived control has a negative correlation with perceived competence.

Studies have modelled trust in HCI more completely by considering the moderating factors that affect its relationship to other variables, e.g. avatar familiarity with the "uncanny valley" effect [57]. To our knowledge, the perceived control and trust relationship has not been likewise studied. We instead look into factors known to affect trust, and investigate them as suspected moderators. Humanness, the unconscious attribution of human characteristics to artificial agents, has been linked to willingness to establish common ground with agents [8], and human-agent trust [14]. Closeness, the sense of social intimacy with another, is a part of therapeutic alliance and may be linked to user engagement [30]. Similarly, usability and trust have been linked across domains [50, 51]. Lastly, Attitude towards AI [52], an emerging concept distinct from technology acceptance [9], but which also likely affects trust in a similar capacity [44].

3 METHODOLOGY

This study investigated the relationship between perceived control and different dimensions of trust in burnout support chatbots, as well as its moderating factors, using a simulated therapy-goal learning chatbot and feedback incorporation. We follow an in-between subject controlled experiment to avoid biasing the user's impression of the bot.

3.1 Participants

Over two weeks, 109 participants were recruited for this experiment using 'snowball' recruitment. The inclusion criteria demanded participants be over 18, currently attending a university, or a recent graduate, and comfortably proficient with the English Language. Of those participants, 35 submitted the study questionnaire for analysis, completing the study. Participants who started the questionnaire but did not submit were ignored in our analysis. Participants were not screened for clinical burnout, but assumed to experience some level of stress in their day-to-day life as students. Our sample pool consisted of 63.33% female participants, 33.33% male, and 3.33% otherwise identifying. Of those, 36.67% were undergraduate students, 30% postgraduates, and 33.33% recent graduates. Only 10% of participants reported their technical skills to be below average on a five-point scale.

3.2 Measures

Two measures form the core of this study; the Human-Computer Trust scale [19] measuring the aforementioned trust dimensions (benevolence, competence, perceived risk, and general trust), and the Yu [64] scale for measuring perceived control. Gulati et al. [19] propose a twelve item, five-point Likert scale measure, while the Yu [64] scale consists of five bipolar items (1: Favours item on the left, 7: Favours item on the right). Items measuring perceived risk in Gulati et al. [19] were excluded from this study. Our suspected moderating factors are measured as follows:

- **Usability**, via the UMUX-LITE [33], which is a two-item, seven-point, Likert scale.
- **Humanness or Mindless Anthropomorphism** from Kim and Sundar [28] measuring the unconscious attribution of human characteristics to artificial agents, such as likability. This scale consists of a four-item, ten-point (*very poorly* to *very well*) survey.
- **Closeness** via the Inclusion-of-the-Other-in-the-Self (IOS) [2] scale, a single-item, pictorial, measure. The IOS depicts seven sets of circles of varying degrees of overlap, correlating to degrees of relationship intimacy.
- **Attitude Towards AI**, via two, five-point Likert scale, items. We borrow these items from the twenty item questionnaire proposed by Schepman and Rodway [52].

3.3 Procedure

First, all participants filled out informed consent forms and demographic data surveys. Next, attitude towards AI systems data was collected. Participants were then randomly split into control and experimental groups, where they had one conversation with the chatbot prototype as it suggested three therapeutic goals to them. The control group was unable to correct the chatbot when it was mistaken, only accept or reject them, while the experimental group was. The bot proposed goals to users in the following pattern: irrelevant goal, most relevant goal, irrelevant goal. Goal relevancy was calculated based on the user's answers to an abbreviated version of the Oldenburg Burnout Inventory [10] and was not learned from conversation as presented to the user. The system thus showed some signs of improvement after it has been corrected, but did not give the impression that it could learn perfectly from one piece of feedback. During the post-test, participants completed all remaining measures and answered short-answer questions regarding their experience. The task lasted 25 – 35 minutes. All data were collected via an online survey hosted on Qualtrics. Study procedures were approved by the TU Delft Human Research Ethics Committee (application number: 2005).

The chatbot was developed specifically for this experiment using the framework Rasa[1] and trained on conversation samples created by the primary author. We note that during data collection, we enacted a change to the chatbot's behaviour. The conversational branches were pruned so that the model classified intention more consistently. Issues with intent misclassification persisted after this change, but seemed to be less severe. Before this change we had recruited 19 participants (Control: 4, Experiment: 15), the remaining participants were recruited afterwards. We expect this change made it more likely for participants to have been able to complete the requisite bot interaction before moving on to the post-test survey, but would not have affected their perception of system usability, since the issue was not eliminated.

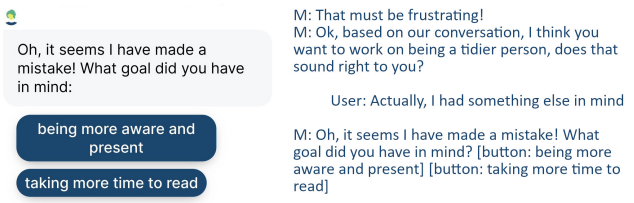

Figure 1: Example of goal correction in the interface (left) and conversation (right)

4 RESULTS

As survey questions did not force participants to answer, some items had missing answers. In those cases, respondent answers on the inventory associated with said missing item were dropped from the analysis.

This analysis was based on Bayesian techniques to enable us to better understand the strength of our evidence. Using Bayesian analysis we were able to perform multimodel inference, report on the likelihood of our models, and, to some extent, adapt to our small sample size [40, 48].

[1] rasa.com

4.1 Does feedback incorporation affect perceived control?

The first step of our analysis was a manipulation check. By comparing perceived control across our experimental and control groups, we verified whether withholding the ability to feedback goals affected participants' sense of control. Users who had trouble completing the conversational flow with the bot, and therefore only saw one or no goal(s) throughout the interaction, were excluded because they did not interact with the goal suggestion feature enough. Of our 35 participants, this excluded 11, for a total of 24 responses, with a further response excluded for missing data. However, it is worth noting that users may have underreported how many goals they saw throughout the conversation, but as we did not log conversations we cannot correct for this. We split our data on experimental group, so that we had 12 participants in control, and 11 in experiment. We then compared the two group's mean perceived control using an independent samples Bayesian t-test with the default prior = 0.707 which returned a Bayes Factor (BF_{10}) of 0.985. With a $BF_{10} = 0.985$, observing the data we do is 0.985 time more likely to occur under our model (where perceived control was higher in the experimental group) than the alternative, null, model. Thus, the relationship was anecdotal [27].

4.2 How does perceived control affect trust?

Since the relationship between the experimental group and perceived control was anecdotal, we did not study differences in trust between the participant groups. Thus, we cannot claim any change in trust levels was a result of feedback incorporation. However, since we are interested in perceived control in general, whether it arises from technical issues, design choices, or our experiment, we should investigate the effects of this variable on trust. Thus, the previously excluded users were reintegrated into the analysis pool from this point onwards. We used Bayesian correlation to investigate the relationship of perceived control with benevolence, competence, and general trust directly, with a prior width of $\frac{1}{3}$ [53]. For prior distribution, we used a Cauchy distribution centred around $d = 0.45$ [53], since based on prior literature [20, 24], we expected to observe a medium-size *Cohen's d* effect. Of our 35 participants, two were excluded for missing data. The correlation analysis showed moderate [27] evidence for an influence of perceived control over benevolence with $r = 0.448$, $BF_{10} = 7.150$. Evidence for this relationship remained in the moderate range across all priors (see Figure 2). The exact effect size in this relationship was fairly uncertain; bound with 95% confidence between $0.082 - 0.630$. Meanwhile, competence ($r = 0.291$, $BF_{10} = 1.232$) and general trust ($r = 0.222$, $BF_{10} = 0.749$) only exhibited anecdotal evidence towards a correlation with perceived control. Thus, our data offers support towards **H1**, but not **H2**.

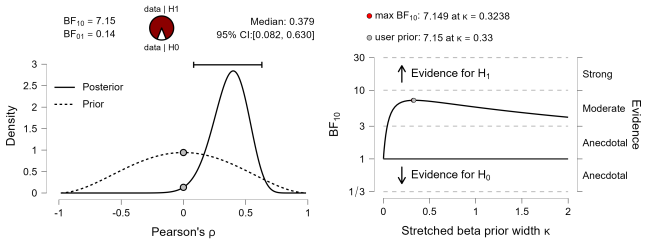

Figure 2: Bayesian Pearson Correlations: Perceived Control & Benevolence

4.3 Do other factors moderate the trust—perceived-control relationship?

To answer **RQ3**, we studied the effects of moderating factors on the relationship between perceived control and each trust dimension using Bayesian linear regression. One such moderation model (benevolence—perceived-control) is seen in Equation 1. This analysis was performed using a beta binomial model prior of $a = 3, b = 3$ [53] [5].

$$\begin{aligned} benevolence = \beta_0 \cdot control + \beta_1 \cdot closeness + \beta_2 \cdot humanness \\ +\beta_3 \cdot attitude + \beta_4 \cdot usuability + \beta_5 \cdot (control * closeness) \\ +\beta_6 \cdot (control * humanness) + \beta_7 \cdot (control * attitude) \\ +\beta_8 \cdot (control * usuability) + \epsilon_i \end{aligned} \quad (1)$$

Usability, humanness, and closeness had little moderating effect on the relationship between perceived control and benevolence, all exhibiting less likelihood to be included in a predictive model of benevolence than perceived control ($BF_{inclusion} < 0.745$). Meanwhile, the interaction effect of perceived control and attitude towards AI exhibited a higher likelihood of inclusion than perceived control alone ($BF_{inclusion} = 1.117 > 0.745$). This interaction was also the only one which was included within the top five predictive models of benevolence (See highlight in Table 1).

Interestingly, the top benevolence model relied on closeness alone as a predictor, with posterior odds of $P(M|Data) = 0.084$. However, the model (highlighted in Table 1) which does contain perceived control (as well as humanness and attitude towards AI), though of lower posterior odds, has 3.093 times the likelihood of co-occurring with our observed data, than a model containing only closeness [60]. If we use frequentist multiple linear regression to compare these two models, we see that the latter achieves higher $R^2(0.645 > 0.361)$ and adjusted $R^2(0.588 > 0.341)$, as well as lower root mean squared error (RMSE) $(0.610 < 0.812)$. Thus, the model containing perceived control, humanness, and attitude towards AI is a better fit over our data [21].

We repeat this analysis with general trust and competence as our dependent variables. No moderating effects emerged. In the case of competence, the best predicting model combined closeness and usability ($P(M) = 0.011, P(M|data) = 0.238, BF_M = 28.524, BF_{10} = 1, R^2 = 0.654$). As for general trust, it was best predicted by combining attitude towards AI and usability ($P(M) = 0.011, P(M|data) = 0.147, BF_M = 15.707, BF_{10} = 1.580, R^2 = 0.586$). The model of competence achieved an $R^2 = 0.591$, an adjusted $R^2 = 0.564$, and an $RMSE = 0.666$. As for trust, the model achieved an $R^2 = 0.570$, an adjusted $R^2 = 0.543$, and an $RMSE = 0.685$.

4.4 How do users feel about providing feedback?

Of our 35 participants, 31 answered the questions on perceived changes in accuracy, and 29 addressed the importance of goal correcting capability. We qualitatively analysed the first third of these responses to create our initial codebook, then refined and finalized it with the remaining responses [16]. Selective coding was used to cluster granular codes.

4.4.1 Did users think their feedback improved the system? Of the 25 participants who indicated whether they felt the chatbot's recommendations improved throughout the conversation when answering this question, 52% said Yes, 44% No, and 4% were Neutral. Thus, more users felt their feedback was effectively incorporated into the chatbot than not, citing factors such as supportiveness, adaptation, and competence (*"Yes, M came up with [the] right conclusions"*). On the other hand, some participants (45.5%) who disagreed also cited competence as a key influence over their perception of model accuracy. Similarly, the remaining neutral responses referenced both competence and adaptation, such that participants felt the system did adapt to their responses but did not prioritize its recommendations well. Participants who did not address the question of accuracy fluctuations in their answers instead discussed technical issues they encountered, including failing to understand user input.

4.4.2 Did users want to provide feedback? In terms of the capacity to edit goals, 76.9% (20 out of 26) of participants wished for, or appreciated, this feature, while 19.2% were not interested, and 3.8% were neutral. Of the participants who were not interested in goal correcting, only one explicitly discusses this feature as *"[...] too much trouble"*. Otherwise, participants desired this opportunity for increased autonomy *"Yes, it would make it more customisable and relevant to me"* and possibly for improved chatbot performance.

5 DISCUSSION

The goal of our first two research questions, **RQ1** and **RQ2**, was to investigate whether providing feedback to an intelligent system would affect a user's perceived benevolence, competence, and general trust. We expected that participants who were able to correct the goals proposed to them by the chatbot, instead of simply accepting or rejecting them, would feel a greater sense of system benevolence (**H1**), but have a lower perception of its competence (**H2**). In **RQ3**, we aimed to model the moderating factors which affect the trust—perceived-control relationship. In this study, our manipulation of perceived control via goal correcting and feedback incorporation was ineffective. Given the sense of confusion among participants in the feedback received, we expect this was due to perceived control being too strongly affected by other aspects of the system, such as intent misclassification, to be manipulated through goal correcting. However, by divorcing the method through which we achieve variance in perceived control from said variance, we were still able to address our hypotheses. The downside in this case is that we are limited in our ability to compare to previous studies we found who, in all but one case (Smith et al. [54]), did not actually measure perceived control. Thus, such studies analyse the

Table 1: Comparison of the 5 best models of the benevolence-perceived control relationship

Models	P(M)	P(M|data)	BF_M	BF_{10}	R^2
Closeness	0.027	0.084	3.302	1.000	0.428
Closeness + Attitude	0.011	0.053	5.089	1.568	0.511
Humanness + Closeness	0.011	0.048	4.595	1.423	0.506
Per. Cont. + Humanness + Attitude + Per. Cont. * Attitude	0.005	0.043	10.006	3.093	0.633
Humanness + Closeness + Attitude	0.006	0.042	7.224	2.240	0.576

relationship between a tangential, but not identical, independent variable (feedback incorporation) and trust [20, 24], where we study perceived control and trust.

5.1 How does perceived control affect trust?

We observed that perceived control was moderately, positively, correlated with benevolence with a Pearson's $r = 0.448$[27], and anecdotally correlated with general trust and competence. This result lends support for hypothesis **H1** that increased control would co-occur with increased benevolence, though not **H2**. One possibility is that the relationship between perceived control and benevolence is the sole reason studies have seen a positive correlation between trust and interactivity in learning systems. While incorporating user feedback into a system may or may not communicate ability on behalf of said system, it does communicate a certain sense of fairness within the human-agent team, similar to that in human teams [29, 61], an element which is closely related to benevolence [22]. On the other hand, perceived control could exhibit a stronger relationship with general trust and competence given a larger dataset. In fact, we expect this would be the case since if only benevolence is tied to control via feedback incorporation, then it does not follow that trust should decrease with feedback as was seen in Honeycutt et al. [24].

5.2 Do other factors moderate the trust—perceived-control relationship?

Attitude towards AI was as a key moderator in our analysis, answering **RQ3**. It was the only factor which influenced benevolence and moderated the relationship between benevolence and perceived control. We can see how such a connection came to be; if a user mistrusts intelligent systems, gestures of benevolence such as feedback solicitation may be viewed as manipulative. If a user is overly positive towards automation, they may place too much emphasis on interactivity and attribute unearned benevolence to the agent, leading to misaligned trust. This is an important finding as attitude towards AI can be understudied in trust modelling compared to acceptance of technology [9].

Surprisingly, closeness emerged as an important predictor of perceived competence. Possibly, users who felt closer to the agent were more likely to overlook mistakes or technical issues during the interaction. Such a relationship is another example of misaligned trust. The user may overestimate the system's capability due to an overly intimate reading of their relationship, or underestimate it because they do not feel close enough. While this is already an issue as over- and under- trust can lead to inappropriate use of intelligent systems [31, 32], it is also an issue because it may encourage developers to maliciously drive up user-agent closeness without improving actual system capability.

5.3 How do users feel about providing feedback?

In both our study and Honeycutt et al. [24], users are confident that their feedback was incorporated into the model. Honeycutt et al. [24] interpret this perception of feedback usage as following from the capacity to provide feedback. We suspect the same, simply allowing the users to reject goals likely impacted their sense of model adaption over time. Similarly, we see a strong desire for feedback incorporation among the users. Though our results indicated goal correcting did not have a strong effect on users' perceived control, it seems that users' nonetheless valued this feature greatly. Since users associated feedback incorporation with concepts of autonomy, and customization, it follows that it should be so valued, and their inclusion would be a justified design strategy [17, 18].

6 LIMITATIONS

Firstly, the small sample size, and limited demographic, restricted our ability to generalize our findings to other populations. When investigating a medium size effect using an in-between subject design frequentist power analysis [13] and Bayesian sample-size literature [15] indicate a population in the range of $154 < n < 216$ would be best, a population larger than recruited in this study. Second, we note the subjectivity in detecting chatbot mistakes in this setting. Though we may have formulated them as irrelevant or relevant, goal recommendations were not always be read thus by the user. Therefore, not every user, even controlling for the number of goals seen, observed the same ratio of hits and misses in their interaction. Since we chose to perform most of our analysis directly on the correlations of perceived control and trust, this variance was not as concerning as it would have otherwise been. Lastly, the non-deterministic nature of our chatbot was an uncontrolled source of mistrust at points. Users had to handle issues of misunderstanding, without researcher involvement, which could have affected their view of the system. As user-agent conversations were not recorded, we could only estimate when these failures occurred based on the answers to our survey.

7 CONCLUSIONS

In this study, we found increased perceived control in burnout support chatbots to positively impact perceived agent benevolence, while perceived competence and general trust in the agent were not likewise affected. We had hoped to study these variables through

the use of feedback incorporation, but effects of this manipulation were non-significant. Though this might have affected our capacity to comment on feedback incorporation and interactivity as a design choice, our users also made clear these features were desired. We also produced multifactor models of different trust components, highlighting attitude towards AI and closeness as two key influences. Through these models, we pinpointed several points in user-agent interactions where misaligned trust could occur. Thus, we propose that the role of perceived control should be a more prominent aspect of future Human-AI interaction studies, and emphasize the need to understand the mechanics of trust in interactive learning more thoroughly.

REFERENCES

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. *Guidelines for human-AI interaction.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[2] Arthur Aron, Elaine N Aron, and Danny Smollan. 1992. Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of personality and social psychology* 63, 4 (1992), 596.

[3] Erin T. Barker, Andrea L. Howard, Rosanne Villemaire-Krajden, and Nancy L. Galambos. 2018. The Rise and Fall of Depressive Symptoms and Academic Stress in Two Samples of University Students. *Journal of Youth and Adolescence* 47 (6 2018), 1252–1266. Issue 6. https://doi.org/10.1007/s10964-018-0822-9

[4] Charles B. Bennett, Camilo J. Ruggero, Anna C. Sever, and Lamia Yanouri. 2020. eHealth to redress psychotherapy access barriers both new and old: A review of reviews and meta-analyses. *Journal of Psychotherapy Integration* 30 (6 2020), 188–207. Issue 2. https://doi.org/10.1037/int0000217

[5] Jeremy C. Biesanz, Carl F. Falk, and Victoria Savalei. 2010. Assessing Mediational Models: Testing and Interval Estimation for Indirect Effects. *Multivariate Behavioral Research* 45, 4 (2010), 661–701. https://doi.org/10.1080/00273171.2010.498292

[6] Pooja Chandrashekar. 2018. Do mental health mobile apps work: evidence and recommendations for designing high-efficacy mental health mobile apps. *mHealth* 4 (3 2018), 6–6. https://doi.org/10.21037/mhealth.2018.03.02

[7] Helen Christensen, Kathleen M Griffiths, and Louise Farrer. 2009. Adherence in Internet Interventions for Anxiety and Depression: Systematic Review. *J Med Internet Res* 11, 2 (24 Apr 2009), e13. https://doi.org/10.2196/jmir.1194

[8] Kevin Corti and Alex Gillespie. 2016. Co-constructing intersubjectivity with artificial conversational agents: People are more likely to initiate repairs of misunderstandings with agents represented as human. *Computers in Human Behavior* 58 (2016), 431–442. https://doi.org/10.1016/j.chb.2015.12.039

[9] Fred D. Davis. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* 13, 3 (1989), 319–340. http://www.jstor.org/stable/249008

[10] Evangelia Demerouti, Evangelia Demerouti, Arnold B. Bakker, Ioanna Vardakou, and Aristotelis Kantas. 2003. The Convergent Validity of Two Burnout Instruments. *European Journal of Psychological Assessment* 19, 1 (2003), 12–23. https://doi.org/10.1027//1015-5759.19.1.12

[11] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64 (3 2018), 1155–1170. Issue 3. https://doi.org/10.1287/mnsc.2016.2643

[12] Gunther Eysenbach. 2005. The Law of Attrition. *J Med Internet Res* 7, 1 (31 Mar 2005), e11. https://doi.org/10.2196/jmir.7.1.e11

[13] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.

[14] Asbjørn Følstad, Cecilie Bertinussen Nordheim, and Cato Alexander Bjørkli. 2018. What makes users trust a chatbot for customer service? An exploratory interview study, In Internet Science. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11193 Lncs, 194–208. https://doi.org/10.1007/978-3-030-01437-7_16

[15] Qianrao Fu, Herbert Hoijtink, and Mirjam Moerbeek. 2021. Sample-size determination for the Bayesian t test and Welch's test using the approximate adjusted fractional Bayes factor. *Behavior Research Methods* 53, 1 (2021), 139–152.

[16] William J. Gibson and Andrew Brown. 2009. *Working with Qualitative Data.* Sage, London.

[17] Google. 2022. Explainability + Trust. https://pair.withgoogle.com/chapter/explainability-trust/

[18] Google. 2022. Feedback + Control. https://pair.withgoogle.com/chapter/feedback-controls/

[19] Siddharth Gulati, Sonia Sousa, and David Lamas. 2019. Design, development and evaluation of a human-computer trust scale. *Behaviour and Information Technology* 38 (10 2019), 1004–1015. Issue 10. https://doi.org/10.1080/0144929x.2019.1656779

[20] Robert S. Gutzwiller and John Reeder. 2021. Dancing With Algorithms: Interaction Creates Greater Preference and Trust in Machine-Learned Behavior. *Human Factors* 63, 5 (2021), 854–867. https://doi.org/10.1177/0018720820903893

[21] Curt Hagquist and Magnus Stenbeck. 1998. Goodness of fit in regression analysis– R 2 and G 2 reconsidered. *Quality and Quantity* 32, 3 (1998), 229–245.

[22] Benjamin E. Hilbig, Isabel Thielmann, Johanna Wührl, and Ingo Zettler. 2015. From Honesty–Humility to fair behavior – Benevolence or a (blind) fairness norm? *Personality and Individual Differences* 80 (2015), 91–95. https://doi.org/10.1016/j.paid.2015.02.017

[23] Robert R. Hoffman, Matthew Johnson, Jeffrey M. Bradshaw, and Al Underbrink. 2013. Trust in automation. *IEEE Intelligent Systems* 28 (2013), 84–88. Issue 1. https://doi.org/10.1109/mis.2013.24

[24] Donald Honeycutt, Mahsan Nourani, and Eric Ragan. 2020. Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. arXiv, Hilversum, The Netherlands, 63–72.

[25] Adam O. Horvath. 2006. The alliance in context: Accomplishments, challenges, and future directions. *Psychotherapy: Theory, Research, Practice, Training* 43 (9 2006), 258–263. Issue 3. https://doi.org/10.1037/0033-3204.43.3.258

[26] Wayne K. Hoy and Megan Tschannen-Moran. 1999. Five Faces of Trust: An Empirical Confirmation in Urban Elementary Schools. *Journal of School Leadership* 9, 3 (1999), 184–208. https://doi.org/10.1177/105268469900900301

[27] Harold Jeffreys. 1998. *The theory of probability.* OUP Oxford, Oxford, United Kingdom.

[28] Youjeong Kim and S. Shyam Sundar. 2012. Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior* 28 (1 2012), 241–250. Issue 1. https://doi.org/10.1016/j.chb.2011.09.006

[29] M. A. Korsgaard, David M. Schweiger, and Harry J. Sapienza. 1995. Building Commitment, Attachment, and Trust in Strategic Decision-Making Teams: The Role of Procedural Justice. *Academy of Management Journal* 38, 1 (02 1995), 60. https://doi.org/10.5465/256728

[30] Tobias Kowatsch, M. Nißen, Dominik Rüegger, and Mirjam Stieger. 2018. The Impact of Interpersonal Closeness cues in Text-based Healthcare Chatbots on Attachment bond and the Desire to continue interacting: an Experimental Design. In *26th European Conference on Information Systems.* ECIS 2018, Portsmouth, UK. http://www.alexandria.unisg.ch/254284/

[31] John D. Lee and Neville Moray. 1994. Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies* 40, 1 (1994), 153–184. https://doi.org/10.1006/ijhc.1994.1007

[32] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[33] James R. Lewis, Brian S. Utesch, and Deborah E. Maher. 2013. UMUX-LITE: When There's No Time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(Chi '13).* Association for Computing Machinery, New York, NY, USA, 2099–2102. https://doi.org/10.1145/2470654.2481287

[34] G. Lietaer, J. Rombauts, R. Van Balen, and G.T. Barrett-Lennard. 1990. *The therapy pathway reformulated.* Leuven University Press, Leuven, Belgium, 123–153.

[35] Sarah Ketchen Lipson, Emily G. Lattie, and Daniel Eisenberg. 2019. Increased rates of mental health service utilization by U.S. College students: 10-year population-level trends (2007-2017). *Psychiatric Services* 70 (1 2019), 60–63. Issue 1. https://doi.org/10.1176/appi.ps.201800332

[36] Edwin A. Locke and Gary P. Latham. 2002. Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist* 57 (9 2002), 705–717. Issue 9. https://doi.org/10.1037/0003-066x.57.9.705

[37] Christina Maslach, Susan Jackson, and Michael Leiter. 1997. *The Maslach Burnout Inventory Manual.* Vol. 3. Consulting Psychologists Press, Palo Alto, California. 191–218 pages.

[38] D. Harrison McKnight, Michelle Carter, Jason Bennett Thatcher, and Paul F. Clay. 2011. Trust in a specific technology: An investigation of its components and measures. *ACM Trans. Manag. Inf. Syst.* 2 (2011), 12:1–12:25. https://doi.org/10.1145/1985347.1985353

[39] Margaret McLafferty, Coral R. Lapsley, Edel Ennis, Cherie Armour, Sam Murphy, Brendan P. Bunting, Anthony J. Bjourson, Elaine K. Murray, and Siobhan M. O'Neill. 2017. Mental health, behavioural problems and treatment seeking among students commencing university in Northern Ireland. *PLOS ONE* 12, 12 (12 2017), 1–14. https://doi.org/10.1371/journal.pone.0188785

[40] Daniel McNeish. 2016. On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal* 23, 5 (2016), 750–773.

[41] David C. Mohr, Pim Cuijpers, and Kenneth Lehman. 2011. Supportive accountability: A model for providing human support to enhance adherence to eHealth interventions. *Journal of Medical Internet Research* 13 (2011), e30. Issue 1. https://doi.org/10.2196/jmir.1602
[42] Michael I Norton, Daniel Mochon, and Dan Ariely. 2012. The IKEA effect: When labor leads to love. *Journal of consumer psychology* 22, 3 (2012), 453–460.
[43] Hanne Weie Oddli, John Mcleod, Sissel Reichelt, and Michael Helge Rønnestad. 2014. Strategies used by experienced therapists to explore client goals in early sessions of psychotherapy. *European Journal of Psychotherapy & Counselling* 16, 3 (2014), 245–266. https://doi.org/10.1080/13642537.2014.927380
[44] Paul A Pavlou. 2003. Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model. *International journal of electronic commerce* 7, 3 (2003), 101–134.
[45] Olga Perski, David Crane, Emma Beard, and Jamie Brown. 2019. Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? An experimental study. *Digital Health* 5 (2019). https://doi.org/10.1177/2055207619880676
[46] Jacqueline Pitchforth, Katie Fahy, Tamsin Ford, Miranda Wolpert, Russell M. Viner, and Dougal S. Hargreaves. 2019. Mental health and well-being trends among children and young people in the UK, 1995-2014: Analysis of repeated cross-sectional national health surveys. *Psychological Medicine* 49 (6 2019), 1275–1285. Issue 8. https://doi.org/10.1017/s0033291718001757
[47] Pnn. 2020. PERSBERICHT: Bijna helft promovendi heeft vergroot risico op mentale klachten, 40% overweegt te stoppen. https://hetpnn.nl/2020/08/26/persbericht-bijna-helft-promovendi-heeft-vergroot-risico-op-mentale-klachten-40-overweegt-te-stoppen/
[48] Daniel S. Quintana and Donald R. Williams. 2018. Bayesian alternatives for common null-hypothesis significance tests in psychiatry: A non-technical guide using JASP. *BMC Psychiatry* 18 (6 2018), 1–8. Issue 1. https://doi.org/10.1186/s12888-018-1761-4
[49] Carl R. Rogers. 1961. *On becoming a person: A therapist's view of psychotherapy.* Houghton Mifflin, New York, NY, US.
[50] Marie Christine Roy, Olivier Dewit, and Benoit A Aubert. 2001. The impact of interface usability on trust in web retailers. *Internet research* 11, 5 (2001), 388–398.
[51] Davide Salanitri, Chrisminder Hare, Simone Borsci, Glyn Lawson, Sarah Sharples, and Brian Water Fi Eld. 2015. Relationship between trust and usability in virtual environments: An ongoing study. In *Human-Computer Interaction: Design and Evaluation (Lecture Notes in Computer Science)*, Masaaki Kurosu (Ed.). Springer, Cham, 49–59. https://doi.org/10.1007/978-3-319-20901-2_5 17th International Conference on Human-Computer Interaction, HCI 2015 ; Conference date: 05-08-2015 Through 07-08-2015.
[52] Astrid Schepman and Paul Rodway. 2020. Initial validation of the general attitudes towards Artificial Intelligence Scale. *Computers in Human Behavior Reports* 1 (2020), 100014. https://doi.org/10.1016/j.chbr.2020.100014
[53] Xenia Schmalz, José B Manresa, and Lei Zhang. 2020. What is a Bayes Factor? https://doi.org/10.31219/osf.io/vgqbt Preprint..
[54] Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2018. Closing the Loop: User-Centered Design and Evaluation of a Human-in-the-Loop Topic Modeling System. In *23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 293–304. https://doi.org/10.1145/3172944.3172965
[55] Matthias Söllner, Axel Hoffmann, Holger Hoffmann, and Jan Marco Leimeister. 2012. How to Use Behavioral Research Insights on Trust for HCI System Design. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems* (Austin, Texas, USA) *(Chi Ea '12)*. Association for Computing Machinery, New York, NY, USA, 1703–1708. https://doi.org/10.1145/2212776.2223696
[56] Changwon Son, Sudeep Hegde, Alec Smith, Xiaomei Wang, and Farzan Sasangohar. 2020. Effects of COVID-19 on college students' mental health in the United States: Interview survey study. *Journal of Medical Internet Research* 22 (9 2020), 14 pages. Issue 9. https://doi.org/10.2196/21279
[57] Stephen Wonchul Song and Mincheol Shin. 2022. Uncanny Valley Effects on Chatbot Trust, Purchase Intention, and Adoption Intention in the Context of E-Commerce: The Moderating Role of Avatar Familiarity. *International Journal of Human–Computer Interaction* 0, 0 (2022), 1–16. https://doi.org/10.1080/10447318.2022.2121038 arXiv:https://doi.org/10.1080/10447318.2022.2121038
[58] Helen M. Stallman. 2010. Psychological distress in university students: A comparison with general population data. *Australian Psychologist* 45, 4 (2010), 249–257. https://doi.org/10.1080/00050067.2010.482109
[59] Stefano Teso and Kristian Kersting. 2019. Explanatory interactive machine learning, In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (Honolulu, HI, USA). *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* 1, 239–245. https://doi.org/10.1145/3306618.3314293
[60] Don van den Bergh, Merlise A Clyde, Akash Raj, Tim de Jong, Quentin F Gronau, Maarten Marsman, Alexander Ly, and Eric-Jan Wagenmakers. 2020. A Tutorial on Bayesian Multi-Model Linear Regression with BAS and JASP. https://doi.org/10.31234/osf.io/pqju6
[61] KEES VAN DEN BOS, RIËL VERMUNT, and HENK A. M. WILKE. 1996. The consistency rule and the voice effect: the influence of expectations on procedural fairness judgements and performance. *European Journal of Social Psychology* 26, 3 (1996), 411–428. https://doi.org/10.1002/(SICI)1099-0992(199605)26:3<411::AID-EJSP766>3.0.CO;2-2
[62] Kai Wang, Deepthi S. Varma, and Mattia Prosperi. 2018. A systematic review of the effectiveness of mobile apps for monitoring and management of mental health symptoms or disorders. *Journal of Psychiatric Research* 107 (12 2018), 73–78. https://doi.org/10.1016/j.jpsychires.2018.10.006
[63] Weiquan Wang and Izak Benbasat. 2005. Trust in and Adoption of Online Recommendation Agents ⋆. *Journal of the Association for Information Systems* 6 (2005), 73. Issue 3.
[64] Guo Yu. 2018. *Effects of timing on users' perceived control when interacting with intelligent systems.* Cambridge University, The Old Schools, Trinity Ln, Cambridge CB2 1TN. https://doi.org/10.17863/cam.37907

A APPENDIX: MODIFIED QUESTIONNAIRE: PERCEIVED CONTROL

This scale was developed to measure control in intelligent-system-user interaction. Where Yu [64] used the consistency of interaction rhythm in mixed initiative chatbots to manipulate sense of control, we used feedback incorporation. Therefore, the statement *I was controlling the pace* was rewritten as *I was controlling M's understanding of me.* Moreover, while the original version used the statement *The software intended to challenge me* to convey feelings of frustration with an inconsistent system, we used *M intended to complete its own task* to better align with the kind of impression a negative interaction with a supposedly adaptive system would invoke.

- How did you feel during the task?
  - M adapted to me (1) - I adapted to M (7).
  - I was controlling M's understanding of me (1) - M was controlling its understanding of me (7).
  - M intended to help me (1) - M intended to complete its own task (7).
  - I felt relaxed during this interaction (1) - I felt stressed during this interaction (7).
  - I felt confident using this system (1) - I felt unconfident using this system (7).

B APPENDIX: SELECTED ITEMS QUESTIONNAIRE: ATTITUDE TOWARDS AI

- How interested are you in using artificially intelligent systems in your daily life? Not interested at all (1) Extremely interested (5)
- How do you feel about the use of artificially intelligent systems becoming more common? Extremely negative (1) Extremely positive (5)

C APPENDIX: OPEN ANSWER QUESTIONNAIRE

- Do you think M's goal recommendations improved over the course of your conversation? Why?
- Did you feel the ability to change the suggested goals was important to you? Why?
- What affected your sense of trust in M the most?

D APPENDIX: CODEBOOK

Table 2: Experiment Study Codebook with Descriptions and Examples

Code	Description	Examples	Category	Count
supportiveness	Includes references to mental health support capacity such as sense of care. Many emotion words are seen in this code (warm, calm, hope, cold, abrupt). Can be high or low.	As a bot it is unable to judge and I don't feel the pressure of talking to a real person	Function	15
competence	Relates to correctness of behaviour and technical capability as a system and a chatbot. Dimensional, can be high or low. Different from ease-of-use-disorientation scale in that it does not focus on technical bugs.	Yes, because the bot does not necessarily understand what I'm saying, so it's very useful to be able to correct it	Function	27
benevolence	Relates to friendliness, good-intention, and care afforded the user by the system. Dimensional, can be high or low. Different from supportiveness as it is not intrinsically tied to system functions.	The responses were professional and friendly	UX	7
ease-of-use	References to speed, directness, ease-of-use, or conciseness as perceived by the user. Not dimensional. Low ease-of-use is in disorientation.	No because [goal correcting]'s too much trouble	Function	1
disorientation	Technical or conceptual difficulties creating unclear expectations for users. Can be in regard to the prototype itself or to the study design. Not dimensional. Low disorientation is in ease-of-use.	She didn't understand my goal at first, so I had to go over the same conversation again	Function	14
autonomy	References to editing, choosing, personalizing or controlling aspects of the systems whether desired, praised, or unhelpful. Can be high or low.	Yes, [goal correcting] gave a freedom of choice, the control is in my hands	UX	14

continues on next page

adaptation	References to changing behaviour on the part of M over the course of the interaction. Dimensional. Can be high (adaptive) or low (failure to adapt).	Yes, it went from completely generic to somewhat personalized	Function	9
attitude to AI	References to users' pre-existing conceptions and expectations of AI. Dimensional. Can be high (positive attitude) or low (negative attitude).	Personally, I do not trust a system like M very much. Not because I consider it to be hostile in some way (though I might if this was provided by an insurance company), but because it is rather transparently a chatbot. I don't generally trust systems like these because they can break easily.	UX	3
appropriateness	Includes references to M's capacity to understand, and respect, relationship boundaries. Similarly, the suitability of the prototype to its purpose and context. Can be low (inappropriate) or high (well suited).	"[…]. Also, when burnt out, the last thing I want to do is add the additional stress of learning a new language."	UX	4
no improvement	The participant explicitly states they perceived no improvement in chatbot recommendation accuracy.	No, I did not understand why things where suggested	Function	11
neutral on improvement	The participant explicitly states they don't know whether the chatbot recommendation accuracy improved.	Hard to say. The goals were definitely taking the stock answers into account, but it weighted the desire to learn new things over feeling unable to cope with studies and suggested learning a new language on top of study pressures. I think it could have prioritised getting on top of studies instead of trying to add something else to the pile	Function	1
improvement	The participant explicitly states they perceived an improvement in chatbot recommendation accuracy.	Yes, M understood my needs better	Function	13

continues on next page

explainability	Relates to the understanding of M's motivations, internal workings, and reasoning. Largely expressed as the desire to understand or the inability to understand.	I dont know what information is in M and what the recommendations are based on.	UX	3
usefulness	References to the usefulness of the prototype and its tools to the user. Can be low (useless) or high (useful)	Asking for another suggestion would have been useful, definitely. [...].	Function	5
neutral on correcting	The participant explicitly states they are neutral on the incorporation of ting in their experience with M.	This may have been helpful just to get a better idea of the scope of goals M has to recommend, rather than going through them one by one, but i personally didn't mind [...].	Function	1
no correcting	The participant explicitly states goal correcting was not or would not be important to their experience with M.	No because [goal correcting]'s too much trouble	Function	5
want correcting	The participant explicitly states goal correcting was or would be important to their experience with M.	Yes, [the goals] would make it more customisable and relevant to me	Function	20
external context	References to the context in which the prototype is used, in this case a research study.	[What most affected my trust in M was that] It has academic support.	UX	2
privacy	Includes explicit references to data privacy concerns by participants.	[...] M itself didn't assure me that our conversation would be private. If we removed the context of this being a study that would definitely affect my sense of trust for it	UX	1