



Can You Trust the Black Box? The Effect of Personality Traits on Trust in AI-Enabled User Interfaces

Martin Böckle^(✉), Kwaku Yeboah-Antwi, and Iana Kouris

BCG Platinion, Design and Engineering, Berlin, Germany

{boeckle.martin,yeboah-antwi.kwaku,kouris.iana}@bcgplatinion.com

Abstract. Human-centred artificial intelligence is a fast-growing research stream within the artificial intelligence (AI) and human–computer interaction (HCI) communities. One key focus of this stream is the enablement of trust between end users and the intelligent solution. Although, the current body of literature discusses and proposes a range of best practices for the design of user interfaces for intelligent solutions, there is a dearth of research how such interfaces are perceived by users and especially focusing on trust in these interfaces. In this paper, we investigate how the Big Five personality traits affect trust in AI-enabled user interfaces. We then experimentally verify which design best practices and guidelines proposed by Google enable trust in AI-enabled user interfaces for the different personality types. Initial results ($n = 211$) reveal that three of the Big Five personality traits – *Extraversion, Agreeableness and Open-Mindedness* – show a significant correlation between the degree of the personality trait and trust in the proposed storyboards. In addition, we identified significant positive relationships between the perception of trust by users and four out of the twelve design principles: *review implicit feedback; connect the feedback to UX changes; create opportunities for feedback; fail gracefully and highlight failure*. This paper is of a highly explorative character and provides first experimental results on designing for trust to the HCI/AI community and also highlights future research directions in the form of a research agenda.

Keywords: Human-centred AI · Personality traits · HCI/AI · Big five

1 Introduction

Human-Centred Artificial Intelligence involves considering human needs when developing intelligent solutions. This approach has attracted a lot of attention in the AI and human–computer interaction (HCI) communities [1]. Although Grudin [3] states that, “both research disciplines have been divided by a common focused goal”, current research endeavours are bridging this gap [4, 5]. They are accomplishing this by resolving relevant issues not only from a technological perspective but also by highlighting that humans still represent the most central and critical element in many scenarios to which machine learning (ML) algorithms are applied [4]. AI has already been used to support

the decision-making process in various application domains [6–8]. Examples include the medical domain [6, 11, 12], where AI has been used to classify computed tomography images to aid in the early diagnosis of retinal diseases [6], or in finance, supporting humans in risk management and option pricing decisions [7, 13]. Furthermore, AI [8, 14] has found applications in healthcare where it assists humans with patient administration, patient monitoring and clinical decisions [8]. Although these examples highlight promising early results of the application of AI to support human decision-making in different domains, the design of intelligent systems that provide seamless interaction for individual end users by understanding their different needs is still an ongoing challenge for human-centred AI. In fact, Xu and Riedl [1, 2] argue that while technological factors are of major importance because technology is developing at such a fast pace, nontechnical factors must also receive considerable attention.

From a HCI perspective, the first two waves of AI can be considered failures because they neglected human needs [2]. The focus was predominantly driven by academia and mainly on technological solutions. The third wave of AI, starting in approximately 2006 and characterized by major breakthroughs in the application of deep learning to big data, pattern and speech recognition [2], has been shaping up differently. Very importantly, there is a conversation on how to combine these technological advances with a human-centred approach. There is a clear understanding that intelligent systems need to be designed while taking into account that they form part of a larger ecosystem consisting of human stakeholders including users, operators and clients. This has been defined as human-centred AI [1].

There is increasing awareness that the AI goals of intelligent interfaces would strongly benefit from the application of user-centred design principles, especially the principles concerning user testing defined in the HCI community [9]. HCI needs to tackle several challenges in the design of AI-enabled interfaces. Firstly, AI algorithms deal with a high level of complexity and therefore the interfaces need to provide transparency and better explanations. Secondly, AI-enabled user interfaces are generally designed for more long-term interactions since algorithms initially need to be “trained up” before obtaining the full benefit of the user-experience (UX) design [9]. Consequently, there is a need for a user-centred approach which helps users understand the characteristics and output of the AI algorithm in order to address the AI black-box problem [2], to enable trust and to tackle issues such as social responsibility, fairness and accountability [1].

If systems with AI components are designed to be human centric and enable trust, Riedl [1] highlight two main aspects for consideration. Firstly, AI and ML systems must be able to produce understandable explanations. This is mainly covered by the emerging and fast-growing research stream of explainable AI (xAI), which provides toolkits and taxonomies for AI explainability techniques [16, 17]. Secondly, human-centred AI and ML systems need to consider the behaviour and objectives of different personalities in order to address their individual needs [15]. The current paper investigates this second aspect and examines how it affects trust in AI-enabled user interfaces by addressing the following research question:

RQ 1: Do personality traits affect trust in human-centred AI-enabled user interfaces?

In the current body of HCI/AI literature, there is consensus around a set of proposed best practices and guidelines for building human-centred AI-enabled user interfaces [18, 19]. If personality traits affect trust in these interfaces, then it is important to understand which best practices and guidelines encourage a high level of trust for different personality traits, an area which considerably lacks research. Such knowledge will be very useful in the design of human-centred AI approaches. This paper therefore attempts to answer a second research question:

RQ 2: Which best practices and guidelines for the design of human-centred AI-enabled interfaces instil the highest levels of trust for different personalities?

As such, this study has two main objectives. First, to investigate the influence of personality traits on trust in AI-enabled user interfaces and second, to experimentally verify which design best practices and guidelines enable trust in AI-enabled user interfaces according to personality type. We also propose a research agenda intended as a basis for future research directions and to guide the practical application of human-centred AI design by considering users' personality traits.

To the best of our knowledge, this is one of the first papers to investigate the relationship between personality traits and trust in AI-enabled user interfaces and therefore contributes new design knowledge to the current body of HCI/AI literature and community.

2 Research Background and Related Work

To identify and develop the research gap and the definition of the proposed research design, we studied several streams within the literature on HCI/AI.

2.1 Human-Centred AI

There are several emerging and fast-growing research streams under the umbrella of human-centred AI that highlight different aspects of intelligent solutions. One stream focuses on ethically responsible AI [2], which aspires to avoid discrimination and achieve fairness. Another stream focuses on designing explainable, useful and useable AI solutions [1, 2], as these features have been neglected in the past. As AI is increasingly applied in multiple end-user applications, most users, especially those with a limited technical background, perceive intelligent systems as a black box. This phenomenon causes end users to ask systems questions such as: "Why did you do that? Why is this the result? Why did you succeed or fail? When can I trust you?" [1, 2, 10]. These questions represent the baseline of explainable AI (xAI), which is currently a popular topic in human-centred AI literature. xAI attempts to address these systems' opacity for the end user by explaining why a certain solution is presented. This approach focuses on understanding and interpreting the AI system output. Recent examples of such work include the classification of explanations of AI solutions for different types of user – developers, AI researchers, domain experts, lay end users – by considering their goals [20]. This is a novel framework that comprises relevant theories on human decision-making, e.g.,

how people should and actually reason in order to inform the xAI techniques [21] and a taxonomy focusing on questions about, “What is explained?” (e.g., data or model), “How it is explained?” (e.g., direct/post-hoc, static/interactive) and “At what level?” (e.g., local/global) [22]. These approaches address the research question, “How to and why trust AI-enabled user interfaces and their output?”

2.2 Trust in AI

When end users overcome the perception of uncertainty or risk, they then start to develop trust in technology. This usually happens after an assessment of the technology’s performance, reliability, safety and security [23]. Generally, trust is a very complex construct [24] and within the domain of human–machine interaction (HMI) it is defined as, “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability” [25]. The work of Ferrario et al. [26] discussed the concept of e-trust when end users interact with e-commerce platforms, group chats and online communities [27]. McKnight et al. [28] investigated the concept of initial trust formation, which explains the preliminary acceptance of technology, in contrast to continuous trust development, where trust needs to be maintained over time. These approaches deal with the question, “How is trust formed when interacting in a digital context?”

2.3 Best Practices and Guidelines for the Design of Intelligent Solutions

This research paper aims to answer the question, “Who trusts AI-enabled user interfaces and which ones do they trust?”, rather than the questions put forward at the end of the last two paragraphs, both having already been discussed extensively in the domain. Jamson et al. [29] showed that AI technology is only effective when it assumes some degree of control for users. To support this effective use of AI technology, interaction with AI-enabled user interfaces should therefore aim to build trust in various ways by considering the factors that influence trustworthiness in such systems, as outlined by Ashoori et al. [30]. Google has created a collection of recommended best practices and guidelines from the UX domain that should be followed in order to achieve a human-centred approach to AI that encourages trust [18]. These include how to introduce the end user to the AI system in a meaningful way and how to set expectations for adaptation and create effective mental models. Other best practices and guidelines show that feedback is crucial to developing trust in AI-enabled user interfaces, as well as explaining how to review, collect and connect implicit and explicit feedback to inform and enhance the user’s experience of the product [18]. Google [18] proposed further guidelines on how to define errors and failure and provide a path forward from failure, since AI capabilities can change over time, leading to errors and failures, and the ability to deal with this gracefully is crucial to trust. As explainability is considered one of the major drivers for increasing trust in these systems, there are also best practices that explain how AI systems work by connecting explanations to the end users’ actions with the AI system

output and workings of the system’s optimisation process [18, 20]. In the current paper, we highlight four sets of best practices and guidelines out of six proposed by Google [18], all of which are widely used and accepted in practice. The four sets of guidelines are:

- *Mental Model*: This set of guidelines concerns the end users’ understanding of how AI systems work and how their interactions affect the interface. Generally, mental models aim to set expectations about functionalities and communication limitations.
- *Explainability and Model Confidence*: These guidelines address how the end user receives an appropriate level of explanation regarding how the system works and its degree of confidence in its output. After developing a clear mental model and awareness of the system’s overall capabilities, these guidelines help end users learn how and when to trust the underlying system. They were originally named “Explainability + Trust” but have been renamed here to avoid confusion.
- *Feedback and Control*: This set of guidelines concerns the design of feedback and control mechanisms that provide a meaningful end-user experience (UX) when suggesting personalized content. These mechanisms can also be used to improve the underlying AI model output.
- *Errors and Graceful failures*: These guidelines help identify and diagnose AI context errors and communicate the way forward. Context errors include false starts, misunderstandings and edge cases that cannot be foreseen within the development process. Google suggests these errors should be seen as opportunities to correct the end user’s mental model, encourage the end user to provide feedback and enhance the overall learning process through experimentation and error resolution processes [18].

We focused on these four sets of guidelines from the total of six sets because only these four applied to the context of our evaluation.

Although these UX best practices and guidelines for the design of AI-enabled user interfaces are widely used in the industry, there is a lack of experimental validation of their efficacy in terms of individual user needs that can be associated with different user/personality types [15, 51]. This is very important, since these different user/personality types are known to have distinct needs which should be taken into consideration when implementing a human-centred design approach.

2.4 User-Type Models

User typologies or personality types have a long history of use in the design of personalised solutions, especially in the domain of learning and persuasion. For instance, Böckle et al. [31, 32] used user/player types to define user-centred design possibilities at the intersection of gamification and persuasive technology. These typologies are also very useful in HCI [33] for defining boundaries to ensure successful interaction with AI-enabled systems and user interfaces [15, 34]. Stachl et al. [35] reported that user/personality types are better predictors of AI application usage than basic demographic parameters [36]. Although several different user type models have been discussed in the literature, in this paper, we focused on the Big Five personality types proposed by Goldberg [36]. These types are described in the Table 1.

Table 1. Big Five factors of personality (BFI-2) [54]

Extraversion (E)	Active, assertive, energetic, enthusiastic, outgoing, talkative
Agreeableness (A)	Appreciative, forgiving, generous, kind, sympathetic, trusting
Conscientiousness (C)	Planula
Negative Emotionality (N)	Anxious, self-pitying, tense, touchy, unstable, worrying
Open-Mindedness (O)	Artistic, curious, imaginative, insightful, original, wide interests

3 Research Design

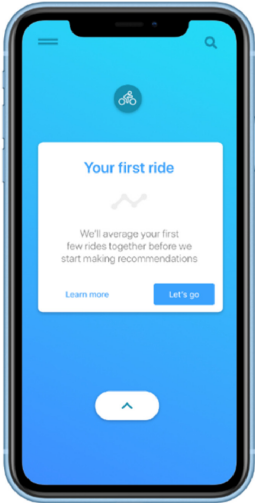
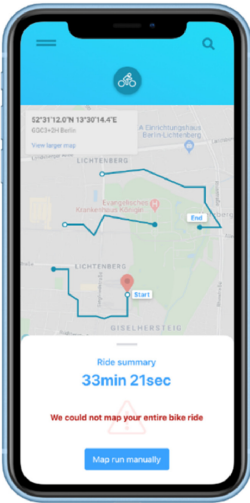
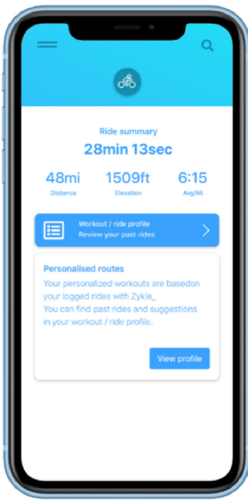
To collect data to study how personality traits affect trust in human-centred AI-enabled user interfaces, we created a survey with storyboards covering the four sets of best practices and guidelines for the design of AI-enabled user interfaces proposed by Google [18] and described in Sect. 2.3. The storyboards show a prototype of a mobile application called “Zycle” created using the sets of best practices from Google [18]. Zycle is an application that helps users who cycle for exercise, a common mode of exercise in Germany, the country of residence of the authors. Zycle has various functionalities such as suggesting routes for rides based on the user’s profile, as well as ride summaries and music playlists. We designed three storyboards for each set of best practices and guidelines, giving a total of twelve storyboards. The storyboards and their matching sets of guidelines are shown in Table 2.

Users were asked to rate each storyboard on the perceived trust. They were given the following scale and asked to pick one option.

- (1) *The system is reliable*
- (2) *I am confident in the system*
- (3) *I can trust the system*
- (4) *I am not suspicious of the system’s intentions, actions or outputs.*

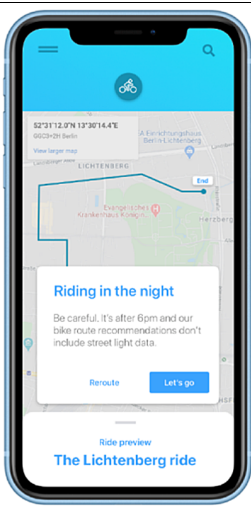
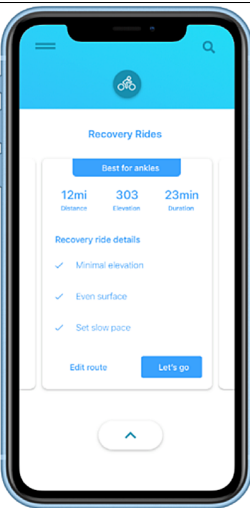
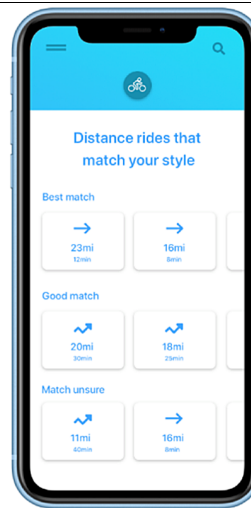
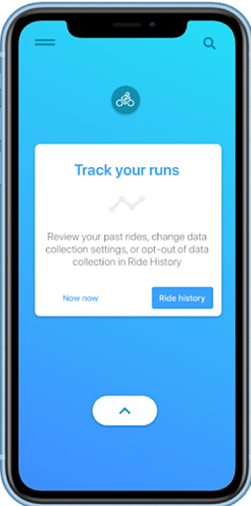
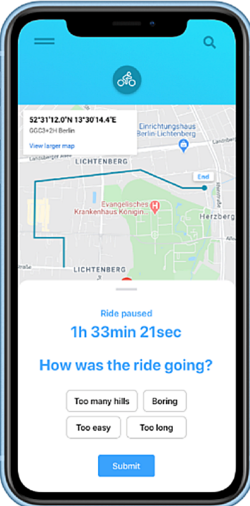
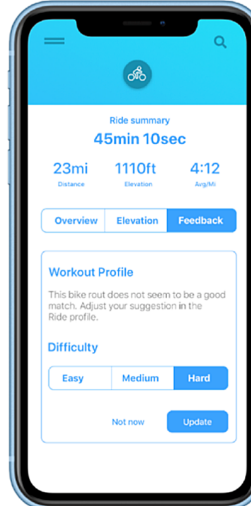
Although the literature discusses several approaches for measuring trust [38, 39], we selected this scale for measuring trust in automated systems [37]. We believe this scale is the most appropriate as it is one of the few that has been empirically tested and used to measure trust in automated systems such as AI-enabled systems.

Table 2. Defined storyboards inspired by [18]

Mental Model Mapping		
<p>S1 Design for experimentation – The interface indicates that the application will average the first few rides together before starting to make recommendations.</p>	<p>S2 Fail gracefully and highlight failure – The interface explains that the application could not map the entire bike ride.</p>	<p>S3 Clearly communicate the limits of AI – The interface highlights the features of the AI component and sets expectations while also helping the end user.</p>
		
Explainability and model confidence		
<p>S1 Articulate data sources – The interface reports a lack of data and suggests that the user uses their own judgement.</p>	<p>S2 Account for situational stakes – The interface explains why certain recommendations have been made in a certain context (e.g., best for ankles).</p>	<p>S3 Communicate model confidence – The interface highlights different recommendations of representing the confidence values as categories (e.g., high/medium/low).</p>

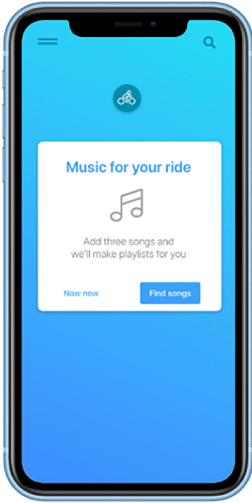
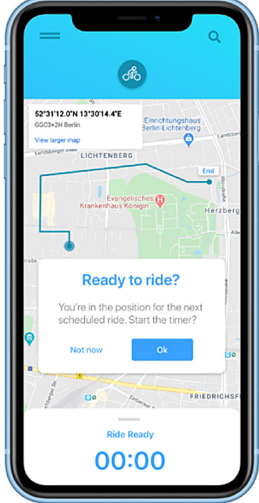
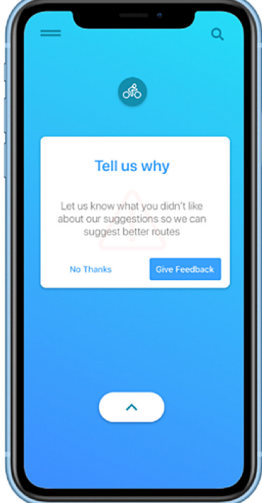
(continued)

Table 2. (continued)

 <p>A mobile app screenshot showing a map of a bike route in Lichtenberg. A white overlay box titled "Riding in the night" contains the text: "Be careful. It's after 6pm and our bike route recommendations don't include street light data." Below the text are "Reroute" and "Let's go" buttons. At the bottom, it says "Ride preview The Lichtenberg ride".</p>	 <p>A mobile app screenshot titled "Recovery Rides". It features a blue button "Best for ankles" and displays ride statistics: "12mi Distance", "303 Elevation", and "23min Duration". Below, under "Recovery ride details", are three checked items: "Minimal elevation", "Even surface", and "Set slow pace". At the bottom are "Edit route" and "Let's go" buttons.</p>	 <p>A mobile app screenshot titled "Distance rides that match your style". It lists "Best match", "Good match", and "Match unsure" categories. Each category shows ride options with distance and duration, such as "23mi 12min" and "16mi 8min".</p>
<p>Feedback and control</p>		
<p>S1 Review implicit feedback – The interface provides the option of looking at past rides and adjusting data collection settings.</p>  <p>A mobile app screenshot titled "Track your runs". It contains the text: "Review your past rides, change data collection settings, or opt-out of data collection in Ride History". At the bottom are "Now now" and "Ride history" buttons.</p>	<p>S2 Connect the feedback to UX changes – The interface allows users to provide feedback, which will have a direct impact on the UX.</p>  <p>A mobile app screenshot titled "Ride paused". It shows a map and the text: "Ride paused 1h 33min 21sec". Below is a survey question "How was the ride going?" with four buttons: "Too many hills", "Boring", "Too easy", and "Too long". A "Submit" button is at the bottom.</p>	<p>S3 Consider user preferences – The interface provides the option to make adjustments if the user's preferences change.</p>  <p>A mobile app screenshot titled "Ride summary". It displays "45min 10sec" and statistics: "23mi Distance", "1110ft Elevation", and "4:12 Avg. Spd". Below are "Overview", "Elevation", and "Feedback" buttons. A "Workout Profile" section contains a message: "This bike route does not seem to be a good match. Adjust your suggestion in the Ride profile." Below is a "Difficulty" section with "Easy", "Medium", and "Hard" buttons, and "Not now" and "Update" buttons at the bottom.</p>

(continued)

Table 2. (continued)

Errors and graceful failures		
<p>S1 Define meaningful error states – Use of error states to inform the user what input the AI needs.</p>	<p>S2 Consider contextual recommendations – The interface offers proactive recommendations.</p>	<p>S3 Create opportunities for feedback - The interface presents the opportunity to receive feedback if the user rejected AI outputs.</p>
		

4 Results

We used the Amazon Mechanical Turk (MTurk) micro-task market to collect the results of the questionnaires. MTurk is a widely accepted and reliable method [40] for collecting end user responses and has also been applied in a variety of peer-reviewed HCI studies [41–45]. Several studies have confirmed the methodological validity of MTurk in various types of survey, including the issues of sampling and participant representatives and potential biases [46–48]. The major advantage of MTurk is that it taps into a diverse population of potential participants who are more easily accessible than via traditional recruitment methods. To obtain high-quality results, we followed the best practices recommended by Goodman and Kurtis by setting a minimum approval rating of 97% and a minimum number of approved tasks of 5,000. To answer our research questions, we applied a multiple regression analysis to the results of the survey to determine which of the predictors, in this case the Big Five personality traits, showed a high correlation with a perception of trust in each of the storyboards.

4.1 Demographic Information

We received 239 complete responses and discarded 28 responses that we considered invalid, giving a total of 211 valid responses. The discarded responses violated certain rules such as completing the questionnaire in less than the minimum time of 15 min. Demographic data revealed that the participants represented a very diverse population with some differences regarding their home country and level of education. The demographic data are shown in Table 3. It shows that 41% of the subjects were female and 59% male. The largest age group (46%) were between 26 and 35 years old and 49% of all respondents had a degree. Most respondents selected India (51%) as their home country, followed by the United States (41%).

Table 3. Demographic data

Total responses (n = 211)	
Gender	Female (41%), Male (59%), Trans (0%), Other (0%)
Age	15–25 (4%), 26–35 (46%), 36–45 (28%), Over 45(22%)
Education	Less than high school (1%), High school (9%), Graduate (16%), College diploma (7.5%), Degree (49%), Master’s degree (15%), Doctoral degree (0), Other (2.5%)
Country	India (51%), USA (41%), Austria (4%), Germany (3%), Singapore (1%)

4.2 Personality Traits and Trust in AI-Enabled User Interfaces

Firstly, we shall address the research question:

RQ 1: Do personality traits affect trust in human-centred AI-enabled user interfaces?

We applied a multiple regression analysis to the results of our survey, focusing on the relationship between the Big Five personality traits and perceived trust in the storyboards and, subsequently, the best practices and guidelines used to design the storyboards. The results are shown in Tables 4, 5, 6 and 7.

Overall, the results reveal that all the user types have differing correlations for forming trust in AI-enabled user interfaces. Three out of five user types, namely those with a high tendency towards *Extraversion*, *Agreeableness* and *Open Mindedness*, show a statistically significant correlation with trust in AI-enabled user interfaces. We could not identify any significant relationships for users who tend towards the user types *Conscientiousness* and *Negative Emotionality*. Our results unequivocally show that personality traits affect trust in human-centred AI-enabled user interfaces.

Our second aim was to answer the research question:

RQ 2: Which best practices and guidelines for the design of human-centred AI-enabled user interfaces instil the highest levels of trust for different personalities?

The following subsections provide a more detailed analysis of the relationship between each Big Five personality trait and trust and attempts to answer RQ1 and RQ2 for each trait.

Extraversion: Individuals with this trait are sociable, gregarious, assertive, talkative, energetic and optimistic [49]. A lower degree of *Extraversion* corresponds to more reserved, independent and quieter users [50]. Our results in Tables 4 and 7 revealed a significant positive relationship between the degree of *Extraversion* and trust in storyboards belonging to the categories *Mental Model* and *Errors and Graceful Failures*. Storyboard S2 on *Mental Model Mapping* highlights the best practice of failing gracefully when the system fails to meet expectations but provides the option to map the run manually. In this case, higher degrees of *Extraversion* correlated positively with an increase in trust when this guideline is followed. Users may see the failure not only as forgivable but something they can help fix, thus increasing their perceived trust in the system. There is a similar positive correlation between trust and higher degrees of *Extraversion* for storyboard S3 on *Errors and Graceful Failures*, which corresponds to the best practice for creating feedback opportunities. These results suggest that extraverted users have a higher degree of perceived trust in systems when they are given chances to supply feedback to fix system failures.

Agreeableness: *Agreeableness* is an important factor in human–computer interactions and is characterised as individuals who are helpful, cooperative, good-natured, sympathetic and tolerant of others [49]. Results in Table 6 show significant positive correlations between the degree of *Agreeableness* and trust in storyboards S1 and S2 for *Feedback and Control*. The first storyboard (S1) deals with the best practice of reviewing implicit feedback and informs end users about an option to opt-out of data collection functionalities. In this context, higher degrees of *Agreeableness* correlated positively with an increase in trust when these guidelines were followed. Generally, since the application of data collection practices may conflict with the level of trust in AI, end users perceive this option positively as a transparent system behaviour. The second storyboard (S2) highlights feedback opportunities by directly influencing the UX of the present ride by asking, “How’s the ride going?”; this reveals the balance between control and automation [18].

In this case, individuals can respond with options such as, “Too many hills”, “Boring” or “Too long”, etc. and the results in Table 6 show that higher degrees of *Agreeableness* correlated positively with an increase in trust when this guideline was applied. Therefore, our results confirm that users have a higher level of trust when they are given chances to opt-out of critical functionalities or providing feedback with a direct impact on the present UX.

Open-Mindedness: *Open-Mindedness* is characterised by intellectual curiosity, a high level of creativity, complex and deep thinking, including the interest in abstract ideas [49]. On the other hand, a low level of *Open-Mindedness* corresponds to conventional, conservative behaviour and a preference for familiar situations [33, 52].

Our results in Tables 4 and 5 show that there is a highly significant inverse relationship between the degree of *Open-Mindedness* and trust for all the storyboards for *Mental Model Mapping* and the third storyboard (S3) for *Explainability and Model Confidence*.

Storyboard S1 in *Mental Model Mapping* demonstrates the best practice of designing for experimentation and reassuring users that these features will not dictate their future experiences [18]. Our results in Table 4 show that higher degrees of *Open-Mindedness* correlate negatively with an increase in trust when this guideline was applied.

Consequently, this shows that these individuals perceive experimentation in which the first few routes will be averaged as untrustworthy. Similarly, when this guideline was followed for the second storyboard (S2), where the route could not be mapped properly, we identified a highly significant negative relationship with increased trust. Users with this personality trait do not tolerate any failures in the system or the option of running the mapping manually. Finally, the last storyboard (S3) aims to communicate the limits of AI by carefully describing the AI features of the system which form the basis of the data used for the proposed recommendations. End users with a higher degree of *Open-Mindedness* did not trust this recommendation. Within this context, Google [18] suggested that there is a risk of integrating a generalised “AI helper”, yet the risk of mistrust is even higher if the system limitations are unclear, as it could lead to over-trust or failing to benefit from the real added value [18]. In the category *Explainability and Model Confidence*, we identified a highly significant negative relationship (Table 5) between the degree of *Open-Mindedness* and the third storyboard (S3), which corresponds to the best practice of how to communicate the model confidence in a meaningful way. Although, the interface breaks down model confidence into three levels – best match, good match, match unsure – this personality type does not trust this type of explanation.

Table 4. Mental model mapping

	Storyboard 1			Storyboard 2			Storyboard 3		
	RC	t Stat	p-value	RC	t Stat	p-value	RC	t Stat	p-value
EXT	0.17	1.62	0.10	0.35*	2.20	0.02	0.24	1.84	0.06
AGR	0.14	0.97	0.33	0.00	0.0	0.99	0.11	0.62	0.53
CON	0.15	1.08	0.28	-0.31	-1.47	0.14	0.03	0.20	0.83
NEG	-0.09	-0.88	0.37	-0.25	-1.47	0.14	-0.11	-0.84	0.40
OPE	-0.29**	-2.43	0.01	-0.57**	-3.05	0.00	-0.52**	-3.47	0.00

Coefficients in bold represent a significant correlation (*p-value ≤ 0.05 , **p-value ≤ 0.01), EXT – Extraversion, AGR – Agreeableness, CON – Conscientiousness, NEG – Negative Emotionality (Neuroticism), OPE – Open-Mindedness, RC – Regression Coefficient)

Table 5. Explainability and model confidence

	Storyboard 1			Storyboard 2			Storyboard 3		
	RC	t Stat	p-value	RC	t Stat	p-value	RC	t Stat	p-value
EXT	0.15	1.2	0.21	0.12	1.12	0.26	0.09	0.81	0.41
AGR	0.13	0.75	0.45	0.16	1.00	0.31	0.10	0.63	0.52
CON	0.14	0.91	0.36	0.18	1.24	0.21	0.19	1.28	0.20
NEG	0.08	0.66	0.50	0.10	0.85	0.39	-0.06	-0.55	0.58
OPE	-0.14	-1.04	0.29	-0.23	-1.75	0.08	-0.31**	-2.37	0.01

Coefficients in bold represent a significant correlation (*p-value ≤ 0.05 , **coefficient p ≤ 0.01), EXT – Extraversion, AGR – Agreeableness, CON – Conscientiousness, NEG – Negative Emotionality (Neuroticism), OPE – Open-Mindedness, RC – Regression Coefficient)

Table 6. Feedback and control

	Storyboard 1			Storyboard 2			Storyboard 3		
	RC	t Stat	P-value	RC	t Stat	P-value	RC	t Stat	P-value
EXT	0.15	1.38	0.16	0.15	1.38	0.16	0.16	1.48	0.13
AGR	0.30*	1.92	0.05	0.30*	1.92	0.05	0.19	1.21	0.22
CON	0.12	0.88	0.37	0.12	0.88	0.37	0.11	0.76	0.44
NEG	0.14	1.22	0.22	0.14	1.22	0.22	0.08	0.68	0.49
OPE	-0.17	-1.36	0.17	-0.17	-1.36	0.17	-0.18	-1.43	0.15

Coefficients in bold represent a significant correlation (*p-value ≤ 0.05 , **p-value ≤ 0.01), EXT – Extraversion, AGR – Agreeableness, CON – Conscientiousness, NEG – Negative Emotionality (Neuroticism), OPE – Open-Mindedness, RC – Regression Coefficient)

Table 7. Errors and graceful failures

	Storyboard 1			Storyboard 2			Storyboard 3		
	RC	t Stat	P-value	RC	t Stat	P-value	RC	t Stat	P-value
EXT	0.12	0.97	0.33	0.09	0.76	0.44	0.24*	2.03	0.04
AGR	0.18	1.06	0.29	0.24	1.37	0.17	0.17	1.01	0.31
CON	0.02	0.12	0.89	0.01	0.06	0.94	-0.01	-0.11	0.91
NEG	-0.06	-0.45	0.64	0.02	0.20	0.83	-0.04	-0.36	0.71
OPE	-0.09	-0.65	0.51	-0.11	-0.74	0.45	-0.01	-0.07	0.94

Coefficients in bold represent a significant correlation (*p-value ≤ 0.05 , **p-value ≤ 0.01), EXT – Extraversion, AGR – Agreeableness, CON – Conscientiousness, NEG – Negative Emotionality (Neuroticism), OPE – Open-Mindedness, RC – Regression Coefficient)

5 Conclusion

The results of the present work aim to support researchers in the design of human-centred AI applications with a major focus on trust. Generally, the application of psychology to HCI can be used to assess design effectiveness, as different personality types perceive and use technology differently. Therefore, we looked into best practices from a UX perspective in order to design AI user interfaces that enable trust. These guidelines indicate how to design intelligent solutions in a meaningful way and cover topics such as *Mental Model Mapping*, *Explainability and Model Confidence*, *Feedback and Control* and *Errors and Graceful Failures*. The storyboards we developed were designed to encourage trust and followed a human-centred approach to AI.

The human-centred AI design elements applied within the prototype captured a holistic picture of a mobile phone cycling application that informs and enhances the users training habits through feedback and suggestions. Regarding the first research question, we identified significant correlations between three of the Big Five personality traits and the proposed storyboards and therefore conclude that personality types do affect trust in

human-centred AI-enabled user interfaces. For people with high levels of *Extraversion* and *Agreeableness*, we found a positive relationship with the aforementioned best practices, while *Open-Mindedness* presented negative relationships. For the analysis of the second research question, we have shown which of the proposed best practices actually instil higher levels of trust according to personality type. We identified positive correlations between personality types and trust for: *review implicit feedback; connect the feedback to UX changes; create opportunities for feedback; fail gracefully and highlight failure.*

We believe that there are multiple means and opportunities to consider different personality traits when designing AI applications, which is an area that has received very little research interest. For instance, the term cognitive compatibility indicates that the structure of the human-machine interface should actually match the user's cognitive style [53]. Recent research also suggests that users actually perform better when they use an interface that has been matched to their personality type [53].

Unfortunately, there is no complete set of design guidelines to describe preferences for specific design elements, as most of the limited number of studies published to date present results for combined personality traits [53]. Consequently, this paper has a highly explorative character and contributes first results to the HCI/AI literature and community by highlighting future research directions in the form of a research agenda.

6 Limitations and Future Research

The current paper is also subject to several limitations. First, we could not identify an appropriate scale for measuring trust in AI-enabled user interfaces within the literature, so we applied certain items from the work of Jian et al. [37]. Second, each individual's perceived trust was measured through responses to screenshots of a prototype design. Particularly in the case of AI applications, trust can change while using the application in everyday life, the scenario in which behavioural data will be generated, which has an impact on the maturity of the AI functions and therefore on the level of trust. Both initial trust and continuing trust development play a major role throughout these stages. Furthermore, certain AI features of the prototype may reveal practical strengths and weaknesses. Third, since we used MTurk we knew little about the participants, although it is already a well-established method for the design of large-scale studies. Yet many unknowns still remain, for instance the cognitive profile of the MTurk workers as well as how prior experience might influence their responses [40]. Fourth, the prototype cycling application and its presentation to the end user may be perceived differently in certain cultures and regions around the world.

While there are many unexplored issues within this domain, as discussed by Alves et al. [15, 53], we believe that this paper makes a valuable contribution to the HCI/AI community by highlighting the importance of trust in AI-enabled interfaces based on different individual personal types. Finally, based on the initial literature review we identified challenges that are summarized in a proposed research agenda in Table 8. This is intended as a research outlook within the emerging and fast-growing research stream of human-centred AI. While the first challenge (C1) focuses on the difficulties of meaningfully integrating personality traits into the design of intelligent solutions, the

second challenge (C2) aims to shed light on how different personality traits actually affect certain types of intelligent support. The third challenge (C3) does not relate to the work of Google and Amershi et al. [18, 19], but instead deals with UI related elements; for instance how to present intelligent support in a significant manner (e.g., buttons, element style, information density, themes, etc.), as discussed in Alves et al. [53]. The fourth challenge (C4) aims to uncover what the modification techniques might look like when attempting to foster initial trust and constantly improve the level of trust and therefore the overall experience of the AI application. Finally, the fifth challenge (C5) investigates how to define reinforcement strategies if the level of trust starts to decline for the end user.

Table 8. Proposed research challenges

C1	To examine and understand the difficulties of incorporating personality traits into the design of intelligent solutions to increase the level of trust. In particular, to consider how they emerge and connect to different AI layers in a meaningful way, in function of changing needs and demands from the end user.
C2	To explore how personality types affect different design elements and tasks of the intelligent solution, for instance the frequency, type and quality of explanations for different types of information.
C3	To examine how individuals perceive as trustful, different elements (e.g. buttons, information density) of a user interface (UI) for an AI system, with a specific focus on trust.
C4	To design and develop meaningful modification techniques for different types of intelligent solution (e.g., personality engines) in order to maintain and constantly improve the present level of trust (initial trust versus continual trust development).
C5	To design for trust reinforcement in AI-enabled interfaces.

References

1. Riedl, M.O.: Human-centered artificial intelligence and machine learning. *Hum. Behav. Emerg. Technol.* **1**(1) (2019)
2. Xu, W.: Toward human-centered ai: a perspective from human-computer interaction. *Interactions* **26**(4), 42–46 (2019)
3. Grudin, J.: AI and HCI: Two fields divided by a common focus. *AI Mag.* **30**, 48–57 (2009)
4. Inkpen, K., Chancellor, S., Choudhury, M.D., Veale, M. and Baumer E.: Where is the human? bridging the gap between AI and HCI. In: *Proceedings of CHI' 19 Extended Abstract*, Glasgow, Scotland, UK (2019)
5. Harper, R.H.: The role of HCI in the Age of AI. *Int. J. Hum.-Comput. Interact.* **35**(15), 1331–1344 (2019)
6. Ting, D., Liu, Y., Burlina, P., Xu, X., Bressler, N.M., Wong, T.Y.: AI for medical imaging goes deep. *Nat. Med.* **24**, 539–540 (2018)
7. Aziz, S., Dowling, M.: Machine learning and ai for risk management. In: Lynn, T., Mooney, J., Rosati, P., Cummins, M. (eds.) *Disrupting Finance*. Palgrave Studies in Digital Business & Enabling Technologies. Palgrave Pivot, Cham (2019)

8. Reddy, S., Fox, J., Purohit, P.M.: Artificial intelligence-enabled healthcare delivery. *J. R. Soc. Med.* **112**, 22–28 (2018)
9. Liebmann, H.: User interface goals. AI opportunities. *AI Mag.* **30**(4), 16–22 (2009)
10. Preece, A.: Asking ‘WHY’ in AI: explainability of intelligent systems – perspectives and challenges. *Intell. Syst. Account. Finan. Manage.* **1**, 1–10 (2018)
11. Montani, S.: Exploring new roles for case-based reasoning in heterogenous AI systems for medical decision support. *Appl. Intell.* **28**, 275–285 (2008)
12. Rastgarpour, M., Shanbehzadeh, J.: Application of AI techniques in medical image segmentation and novel categorization of available methods and tools. In: *Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS)*, vol. 1, Hong Kong (2011)
13. Culkin, R., Das, S.R.: Machine learning in finance: the case of deep learning for option pricing. *J. Invest. Manage.* **15**, 4 (2017)
14. Lysaght, T., Lim, H.Y., Xafis, V., Ngiam, K.Y.: AI-assisted decision-making in healthcare. *Asian Bioeth. Rev.* **11**, 299–314 (2019)
15. Völkel, S.T., Schödel, R., Hussmann, H.: Designing for personality in autonomous vehicles: considering individual’s trust attitude and interaction behavior. In: *Proceedings of CHI Workshop – Interacting with Autonomous Vehicles: Learning from other Domains* (2018)
16. Liao, Q.V., Gruen, D., Miller, S.: Questioning the AI: informing design practices for explainable AI user experiences. In: *Proceedings of CHI 2020, Honolulu, USA* (2020)
17. El-Essady, M., et al.: Towards XAI: structuring the processes of explanations. In: *Proceedings of HCML Workshop at CHI’19, Glasgow, UK* (2019)
18. Google PAIR. People + AI Guidebook (2019). <https://pair.withgoogle.com/guidebook/>
19. Amershi, S., et al.: Guidelines for human-AI interaction. In: *Proceedings of CHI 2019, Glasgow, Scotland, UK* (2019)
20. Ribera, M., Lapedriza, A.: Can we do better explanations? A proposal of User-Centered Explainable AI. In: *Joint Proceedings of the ACM IUI 2019 Workshop, Los Angeles, USA* (2019)
21. Wang, D., Yang, Q., Lim B.: Designing theory-driven user-centric explainable AI. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, Scotland, UK* (2019)
22. Arya, V, et al.: One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. [arXiv:1909.03012](https://arxiv.org/abs/1909.03012) (2019)
23. Arnold, M., et al.: FactSheets: Increasing trust in AI services through supplier’s declarations of conformity. *IBM J. Res. Dev.* **63**, 4/5, 6–13 (2019)
24. Culley, K.E., Madhavan, P.: Trust in automation and automation designers: Implications for HCI and HMI. *Comput. Hum. Behav.* **29**, 2208–2210 (2013)
25. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. *Hum. Fact.* **46**, 50–80 (2004)
26. Ferrario, A., Loi, M., Vigano, E.: In AI we trust incrementally: a multi-layer model of trust to analyze human-artificial intelligence interactions. *Philos. Technol.* **35**, 523–539 (2020)
27. Taddeo, M., Floridi, L.: The case of e-trust. *Ethics Inform. Technol.* **13**(1), 1–3 (2011)
28. McKnight, D.H., Choudhury, V., Kacmar, C.: The impact of initial consumer trust on intentions to transact with a web site: a trust building model. *J. Strateg. Inform. Syst.* **11**(3), 297–323 (2002)
29. Jamson, A.H., Merat, N., Carsten, O.M., Lai, F.C.: Behavioural changes in drivers experiencing highly-automated vehicle control in varying traffic conditions. *Transp. Res. Part C: Emerg. Technol.* **30**, 116–125 (2013)
30. Ashoori, M., Weisz J.D.: In AI we trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes. [arXiv:1912.02675](https://arxiv.org/abs/1912.02675) (2019)

31. Böckle, M., Yeboah-Antwi, K.: Designing at the intersection of gamification and persuasive technology to incentivize energy-saving. In: Pappas I., Mikalef P., Dwivedi Y., Jaccheri L., Krogstie J., Mäntymäki M. (eds.) *Digital Transformation for a Sustainable Society in the 21st Century. I3E 2019. Lecture Notes in Computer Science*, vol. 11701 (2019)
32. Böckle, M., Novak, J., Bick, M.: Exploring gamified persuasive system design for energy saving. *J. Enterp. Inform. Manage.* **33**(6), 1337–1356 (2020)
33. Pillis, E., Green, D.: Personality influences trust differently in virtual and face-to-face teams. *Int. J. Hum. Resour. Dev. Manage.* **9** (2009)
34. Zhou, X., Mark, G., Li, J., Yang, H.: Trusting virtual agents: the effect of personality. *ACM Trans. Interact. Intell. Syst.* **9**, 2–3, Article 10 (2019)
35. Stachl, C., Hilbert, S., Au, J., Buschek, D.De, Luca, A., Bischl, B., Hussmann, H., Bühner, M.: Personality, traits predict smartphone usage. *Euro. J. Pers.* **31**(6), 701–722 (2017)
36. Goldberg, L.R.: An alternative description of personality: the Big-Five factor structure. *J. Pers. Soc. Psychol.* **59**, 1216–1229 (1990)
37. Jian, J., Bisnatz, A., Drury, C.: Foundations for an empirically determined scale of trust in automated systems. *Int. J. Cogn. Ergon.* **4**(1), 53–72 (2000)
38. Schaefer, K.E.: Measuring trust in human robot interactions: development of the “trust perception scale-HRI”. In: Mittu, R., Sofge, D., Wagner, A., Lawless, W. (eds.) *Robust Intelligence and Trust in Autonomous Systems*. Springer, Boston, MA (2016)
39. Madsen, M., Gregor, S.: Measuring human-computer trust. In: *Proceedings of the 11th Australasian Conference on Information Systems*, pp. 6–8 (2000)
40. Paolacci, G., Chandler, J.: Inside the Turk: understanding Mechanical Turk as a participant pool. *Curr. Dir. Psychol. Sci.* **23**(3), 184–188 (2014)
41. Orji, R., Tondello, G.F., Nacke, L.: Personalizing persuasive strategies in gameful systems to gamification user types. In: *Proceedings of CHI 2018*, Montreal, Canada (2018)
42. Toomin, M., Kriplean, T., Portner, C. and Landay, A.J.: Utility of human-computer interactions: toward a science of preference measurement. In *Proceedings of CHI 2011*, Vancouver, Canada (2011)
43. Attali, Y., Arieli-Attali, M.: Gamification in assessment: do points affect test performance? *Comput. Educ.* **83**, 57–63 (2015)
44. Feller, J., Gleasure, R., Treacy, S.: Information sharing and user behavior in internet enabled peer-to-peer lending systems: an empirical study. *J. Inform. Technol.* **32**, 127–146 (2017)
45. Huang, Y., Li, C., Wu, J., Lin, Z.: Online customer reviews and consumer evaluation: the role of review font. *Inform. Manage.* **55**, 430–440 (2018)
46. Bartneck, C., Duenser, A., Moltchanova, E., Zawieska, K.: Comparing the similarity of responses received from studies in Amazon’s Mechanical Turk to studies conducted online and with direct recruitment. *PloS One* **10**(4) (2015)
47. Clifford, S., Jewell, R.M., Waggoner, P.D.: Are samples drawn from Mechanical Turk valid for research on political ideology? *Res. Polit.* **2**(4) (2015)
48. Heen, M.S., Lieberman, J.D., Miethel, T.D.: A comparison of different online sampling approaches for generating national samples. *Center Crime Justice Policy*, **1**, 1–8 (2014)
49. Soto, C.J., John, O.P.: The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *J. Pers. Soc. Psychol.* **113**, 117–143 (2017)
50. Pillis, E., Green, D.: Personality influences trust differently in virtual and face-to-face teams. *Int. J. Hum. Res. Dev. Manage.* **9** (2009)
51. Zhou, J., Luo, S., Chen, F.: Effects of personality traits on user trust in human-machine collaborations. *Journal of Multimodal User Interfaces* **14**, 387–400 (2020)
52. Bruck, C.S., Allen, T.D.: The relationship between Big Five personality traits, negative affectivity, type a behavior, and work-family conflict. *J. Vocat. Behav.* **63**, 457–472 (2003)

53. Alves, T., Natlio, J., Henriques-Calado, J., Gama, S.: Incorporating personality in user-interface design: a review. *Pers. Individ. Differ.* **155** (2020)
54. McCrae, R.R., John, O.P.: An introduction to the five-factor model and its applications. *J. Pers.* **60**, 175–215 (1992)