User Modeling and User-Adapted Interaction (2023) 33:121–157
https://doi.org/10.1007/s11257-022-09336-9

Eye-tracking-based personality prediction with recommendation interfaces

Li Chen[1] · Wanling Cai[1] · Dongning Yan[2] · Shlomo Berkovsky[3]

Received: 12 April 2021 / Accepted in revised form: 23 April 2022 / Published online: 24 June 2022

Abstract

Recent research in behavioral decision making demonstrates the advantages of using eye-tracking to surface insights into users' underlying cognitive processes. Personality, according to psychology definition, accounts for individual differences in our enduring emotional, interpersonal, experiential, attitudinal, and motivational styles. In recommender systems (RS), it has been found that user personality is related to their preferences and behavior, which attracted an increasing attention to the ways to leverage personality into the recommendation process. However, accurate acquisition of a user's personality is still a challenging issue. In this work, we investigate the possibility of automatically detecting personality from users' eye movements when interacting with a recommendation interface. Specifically, we report an experiment that harnesses two recommendation interfaces to collect eye-movement data in several product domains and then utilize the data to predict the users' Big-Five personality traits through various machine learning methods. The results show that AdaBoost combined with Gini index score-based feature selector predicts the traits most accurately, and interface- and domain-specific data allow to improve the accuracy of personality trait predictions. Our findings could inform personality-based RS by improving the process of indirect user personality acquisition.

✉ Li Chen
lichen@comp.hkbu.edu.hk

Wanling Cai
cswlcai@comp.hkbu.edu.hk

Dongning Yan
yandongning@sdu.edu.cn

Shlomo Berkovsky
shlomo.berkovsky@mq.edu.au

[1] Department of Computer Science, Hong Kong Baptist University, Hong Kong, China

[2] School of Mechanical Engineering, Shandong University, Jinan, Shandong, China

[3] Centre for Health Informatics, Macquarie University, Sydney, Australia

Keywords Recommendation interface · Eye-tracking-based personality prediction · Personality-based recommender systems

1 Introduction

Online recommender systems (RSs) have become widely popular over the last decade, since they can effectively reduce online information overload and provide personalized services that assist users in decision making (Ricci et al. 2015). Recently, an increasing attention has been paid to incorporating user personality into the recommendation generation process (Tkalcic and Chen 2015; Tkalcic et al. 2016; Nguyen et al. 2018; Wu et al. 2018), as it has been found that *personality* can be strongly related to users' preferences for items and their interaction with a RS (Rentfrow and Gosling 2003; Hu and Pu 2013; Cantador et al. 2013; Manolios et al. 2019). Motivated by these findings, various approaches have been proposed to develop *personality-based* (or called personality-aware) RS, which are used to address the cold-start issue (Tkalcic et al. 2009; Elahi et al. 2013; Fernández-Tobías et al. 2016) or improve the diversity of the generated recommendations (Ferwerda et al. 2016; Wu et al. 2018). However, the task of accurately acquiring a user's personality for building such a RS largely remains a challenge. In earlier works, explicit approaches via psychological instruments, such as the International Personality Item Pool (IPIP) questionnaire (50 or 100 items) (Goldberg et al. 2006), the Big-Five Inventory (BFI) (44 items) (John et al. 1999), and the Ten Item Personality Inventory (TIPI) (10 items) (Gosling et al. 2003), have mainly been deployed to assess a user's personality. However, as any self-reported instrument, personality inventories are prone to manipulation and faking (Anglim et al. 2018; Fahey 2018), especially in high-stake situations (Ziegler et al. 2011). Hence, they may place burden on users, while their replicability in practice is limited (Tkalcic and Chen 2015). Hence, implicit techniques have been developed in recent years, with the aim of offering an objective and unobtrusive way of acquiring user personality. These methods primarily relied on user-generated social media content (e.g., on Facebook, Twitter, Weibo, or Instagram) (Quercia et al. 2011; Kosinski et al. 2013; Gao et al. 2013; Ferwerda et al. 2015), which were hard to generalize to a broader set of RS. query Please check the edit made in the article title.

In this work, we focus on investigating the possibility of inferring user personality from their observable eye movements on the recommendation interface, which might provide a more feasible and objective approach to identifying user personality for recommender systems. In the field of behavioral decision making, eye-tracking has been demonstrated as a useful tool for accurately capturing users' cognitive processes associated with decision making (Franco-Watkins and Johnson 2011; Glaholt and Reingold 2011; Ashby et al. 2016). It has been found that eye movements can more precisely disclose how users make a decision than classical self-report methods or those based on cursor movements, because of the direct measure of users' visual attention (Franco-Watkins and Johnson 2011; Rojas et al. 2020). Moreover, eye movements provide a rich source of data that allows researchers to understand how users attend to and use information in the construction of their preferences (Cavanagh 2014). For

instance, it has been shown that fixation is an indicator of liking, i.e., users fixate on items they are more likely to select (Stewart et al. 2016; Mitsuda and Glaholt 2014).

Given that the eye movements have been recognized as a valid proxy for users' decision-making processes, it may sound promising to *harness users' eye movement data to detect their personality*, so as to benefit personality-based RS. The increasing sophistication, accessibility, and accuracy of eye-tracking technologies make this idea practical and feasible (Zhang et al. 2017; Valtakari et al. 2020). With this objective in mind, we conducted an experiment, where the eye movements of 130 subjects over their first encountered recommendation interface were recorded, which resulted in a collection of 14,259 gaze data points from 108 valid subjects. We used this data to extract 86 eye-movement features and fed them into various machine learning methods to predict subjects' personality. We adopted the well-studied Big-Five personality model, widely used in RS (Tkalcic and Chen 2015; Tkalcic et al. 2016), which consists of five factors (or called traits): Openness to experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (McCrae and John 1992).

In addition to detecting user personality from observable eye movements, we were also interested to uncover what kind of recommendation interface might be more effective in this regard. To this end, we evaluated two representative RS interfaces and three domains of recommended items. To the best of our knowledge, our work is the first attempt to harness eye-tracking as an implicit source of data for user personality acquisition in RS specifically. We believe that our results could facilitate more effective future personality-based RS, which may rely on user personality detected from their eye movements on the first interface to improve the system's recommendations in the subsequent interaction cycles.

In summary, our work makes three key contributions: (1) we demonstrate the *feasibility of detecting user personality* from their eye movements when interacting with a recommendation interface; (2) we study the *effects of interface design and product domain* on the accuracy of personality detection; and (3) we identify a set of *predictive eye-tracking features and informative components of the interface*.

The rest of this paper is structured as follows: we first introduce in Sect. 2 related work on personality acquisition in RS and eye-tracking research in behavioral decision making. Then, in Sect. 3 we present our experimental setup for collecting eye-movement data, including interfaces, participants, experimental procedure, and eye-tracking features. The results of personality predictions are provided in Sect. 4. Finally, we discuss the practical implications and limitations of the work in Sect. 5 and draw the conclusion in Sect. 6.

2 Related work

2.1 Personality acquisition in recommender systems

Personality, as defined in psychology, explains the key dimensions, in which individuals differ in their enduring emotional, interpersonal, experiential, attitudinal, and motivational styles (McCrae and John 1992; Ajzen 2005). Studies of behavioral decision making found that individual differences, as accounted for by personality, can

take a critical role in affecting how individuals make a decision, e.g., with respect to their risk-taking propensity and preference for intuition or deliberation decision style (Pachur and Spaar 2015; Nicholson et al. 2005). In the area of RS, which may be considered as decision support tools, it was found that personality relates to user preferences for categories of items, e.g., movies, TV shows, music, books, and interest groups in the social media (Rentfrow and Gosling 2003; Cantador et al. 2013; Manolios et al. 2019; Wu et al. 2018). Moreover, personality also influences user behavior when using a RS, such as retention, activity level, and rating patterns (Hu and Pu 2013; Karumur et al. 2018).

Motivated by these findings, a stream of work has recently attempted to incorporate personality into the process of improving the quality of recommendations, in the so-called *personality-based RS* (Tkalcic and Chen 2015; Tkalcic et al. 2016). For instance, in Tkalcic et al. (2009), the authors used personality to improve the nearest-neighbor measure in a collaborative filtering (CF) system and demonstrated that a personality-based similarity measure is more accurate than traditional rating-based measures, particularly in a cold-start situation. In Hu and Pu (2011), the authors developed a cascade hybrid CF, which adopted pure personality-based algorithm to make initial predictions and then applied CF to the user-item matrix. Their evaluation showed that the hybrid method significantly outperformed the traditional approach in sparse datasets.

More recently, Fernández-Tobías et al. (2016) developed three approaches to mitigating the new user problem, respectively, based on personality-based matrix factorization (MF), personality-based active learning, and personality-based cross-domain recommendation. They found that all of the personality-based methods improved performance in real-life datasets, while the personality-based cross-domain method performed the best. Personality has also been incorporated into preference-based RS. For instance, Hu and Pu (2010a) established a personality-based interest profile for each user, which reflected the relationship between personality and user preferences for music genres (Rentfrow and Gosling 2003). Items that best matched the user's profile were then recommended.

Furthermore, some works have taken personality into account for improving the recommendation diversity, as they found that the preferred diversity level of a set of recommendations could be affected by the user's personality traits, such as *Openness to experience* and *Conscientiousness* (Tintarev et al. 2013; Ferwerda et al. 2016; Chen et al. 2013c). For instance, the authors of Chen et al. (2013c) proposed a generalized, dynamic personality-based greedy re-ranking approach for generating diverse recommendations (Wu et al. 2018). Their evaluation demonstrated that this approach was significantly more effective than both non-diversity-oriented and related diversity-oriented methods in terms of the recommendation accuracy and personalized diversity degree, especially in a cold-start setting.

Most of the above personality-based RS rely on psychological instruments to acquire the personality traits, such as the International Personality Item Pool (IPIP) (50 items) in Tkalcic et al. (2009), the Ten Item Personality Inventory (TIPI) in Hu and Pu (2010b), Elahi et al. (2013), Tiwari et al. (2020), and NEO IPIP (20 items) in Tintarev et al. (2013), Cantador et al. (2013). To the best of our knowledge, few works have automatically detected user personality. For example, in Wu and Chen (2015),

the authors inferred users' personality traits from their historical rating data and then incorporated these into a CF recommender.

In another stream of work, we observe a number of studies focusing on the accuracy of personality predictions relying on, e.g., users' micro-blog posts (Quercia et al. 2011), social network activity (Quercia et al. 2011; Kosinski et al. 2013; Gao et al. 2013), game playing behavior (Van Lankveld et al. 2011), mobile phone usage (Chittaranjan et al. 2011), and emails (Shen et al. 2013). That said, the above works are not centered on RS and it remains unclear how such personality predictions could be deployed in RS.

In this work, we set out to investigate the possibility of detecting user personality from their eye movements on a typical recommendation interface. This way, the predictions can be directly used by RS to improve recommendations in the subsequent user interactions. Alternatively, if the system has already obtained user information, the available eye-movement data could complement this information, to facilitate the provision of more accurate recommendations.

2.2 Eye-tracking in behavioral decision making

Recent advancements in eye-tracking technologies have facilitated the use of eye-trackers to provide direct measures of visual attention in reading and information-processing tasks (Rayner 1998). In particular, it has been recognized that the ease of use and affordability of eye-tracking equipment offer "unique and relatively unhindered insights into perceptual, cognitive, motivational, and/or affective processes underlying human behavior" (Ashby et al. 2016). In prior literature, various eye-movement metrics, e.g., saccades, eye fixations, and pupils, have been used to measure information uptake processes, e.g., what information is processed and examined (Dumais et al. 2010; Raptis et al. 2017; Chen et al. 2013a).

In the area of behavioral decision making, eye-tracker showed its promise as a useful process-tracking tool capturing users' cognitive processes (Glaholt and Reingold 2011; Franco-Watkins and Johnson 2011). For instance, it provided deeper insights into the cognitive processes underlying behavioral phenomena, such as the endowment effect (Ashby 2015) and user preferences in risky choices (Stewart et al. 2016). It also highlighted the importance of attention in preference construction, e.g., the difference between the liking and disliking decisions (Mitsuda and Glaholt 2014), and the effectiveness of eye-movements to infer individual decision strategies (Glöckner and Herbold 2011). Rojas et al. (2020) showed that eye movements were more effective than traditional self-reporting methods in measuring the effort needed to make a decision. Also, Franco-Watkins and Johnson (2011) discovered that eye-tracking provided a more reliable measure of attentional processing during decision making than mouse tracking.

The roles of different eye-tracking variables have also been investigated. In early works, fixation and visit variables were mainly considered. For instance, it was found that fixations on an option could indicate a relative preference, such that receiving fixations was more likely to lead to selection (Mitsuda and Glaholt 2014; Stewart et al. 2016). Later on, additional variables have been studied. For example, it was shown

that pupil dilation was predictive of a perceived decision difficulty and mental effort (Kret 2019). Also, measurable parameters of eye movements were extensively used to detect conscious and unconscious activities. Complex features, like gaze pattern and scan path, were found to be reliable indicators of cognitive strategies and attention (Dumais et al. 2010; Raptis et al. 2017). Pupillary response was used as an indicator of cognitive load (Xu et al. 2011; Chen et al. 2013b), and saccade amplitude and fixation duration were used for lie detection (Lim et al. 2013).

Hence, it was suggested that the eye-movement data should be analyzed in a holistic manner, allowing to combine multiple sources of variables to reveal attentional processes associated with information acquisition and use (Franco-Watkins and Johnson 2011). More recent studies attempted to identify the factors of individual differences in visual attention (Ashby et al. 2016). For instance, Rojas et al. (2020) employed eye-tracker as an implicit tool to analyze children's preferences for toys and found that their choices were influenced by stimuli design dimensions, while gender could explain the differences in fixation and visit times.

Although it has been shown that personality could cause differences in the way users process information, little work has focused on investigating the relationship between personality and users' eye-movement behavior within a decision-making framework. For instance, Wilbers et al. (2015) showed that *Extroversion* was negatively correlated with the duration of fixation when people viewed a set of images, independently of the stimulus type, e.g., color, gist, or valence. Likewise, Rauthmann et al. (2012) found that individuals with a higher *Openness to experience* manifested longer fixation duration and dwelling times when viewing abstract animations without any semantic or topical stimulus information, while those with higher *Extroversion* manifested shorter dwelling times. In the area of recommender systems, recent work has studied how personality influences the way users perceive and process explanations (Millecamp et al. 2020, 2021). For example, it was found that users with low *Openness to experience* benefit more from explanations (Millecamp et al. 2020).

There have also been several attempts to detect users' personality traits from their eye-movement data. For example, Hoppe et al. (2018) leveraged eye movements recorded during an everyday task of walking around a university campus to predict users' Big-Five personality traits and perceptual curiosity. Recently, Berkovsky et al. (2019) and Taib et al. (2020) aimed to predict the values of 16 personality traits across different models (such as D3, BIS/BAS, HEXACO models) by leveraging users' eye-movement data as physiological responses to affective image and video stimuli. It was found that personality traits, especially those associated with affect, can be reliably detected in such a setting, while those related to user cognition and behaviors were detected with a lower predictive accuracy.

In this work, we mainly focus on detecting users' Big-Five personality traits from their observable eye movements when they interact with a standard RS interface. Such an interface typically lists a number of options and the user makes a decision whether to accept the recommended option. Given that (i) personality was shown to be related to user preferences (Tkalcic and Chen 2015; Tkalcic et al. 2016) and (ii) eye movements were recognized as a valid proxy for decision making processes (Glaholt and Reingold 2011; Franco-Watkins and Johnson 2011; Ashby et al. 2016), direct

detection of personality from eye-movement data may offer an innovative method for personality acquisition in RS.

3 Experimental setup for data collection

For data collection purpose, we conducted a lab-controlled eye-tracking experiment. We used the Tobii Pro X3-120 Eye Tracker to record users' eye movements when interacting with the recommendation interface. This tracker is slim and compact enough to facilitate unobtrusive data collection, with high accuracy and sampling rate[1]. We set out to utilize users' eye-movement data captured in their first encountered recommendation interface to predict their personality traits. The values of the traits could be useful for the RS to build personality models for new users and then generate personality-aware recommendations in subsequent interactions. We adopted the experimental setup of Chen et al. (2019), Pu and Chen (2007), which showed a list of items as recommendations for users to examine.

In the rest of this section, we present our experimental setting. The studied RS interfaces are outlined in Sect. 3.1, and the predicted personality traits are discussed in Sect. 3.2. Experimental procedure and descriptive statistics of the participants are detailed in Sect. 3.3. We present the features extracted from the captured eye-tracking data in Sect. 3.4. Finally, we discuss the personality prediction machine learning mechanics in Sect. 3.5.

3.1 Recommendation interfaces

We deployed two representative recommendation interfaces, designed to trigger users' decision-making behavior and obtain their eye-movement data. One is the traditional *LIST interface* (Pu and Chen 2006, 2007), where all the recommended items are displayed sequentially, ranked according to their overall popularity (non-personalized) or to the level of match to the user's preferences (personalized), as shown in Fig. 1a. The other is called the *ORG interface*, as originally proposed by Pu and Chen (2006) and later improved by Chen et al. (2019) for organizing recommendations in a structured category view (see Fig. 1b). Specifically, in ORG, all the items, except the top candidate (the first ranked item), are divided into several categories, where both the similarity of items within a category and the dissimilarity of items across different categories are maximized (Chen et al. 2019). Each category is accompanied by a title to explain the similar properties of its contained items, e.g., "*have a better screen size and better opinions on battery and performance, but worse value at price*" for smartphones. Prior studies showed that ORG was more effective than LIST in terms of instilling users' trust and intention to return, since users perceived it to be more competent in aiding them in product comparison, owing to the category structure (Pu and Chen 2006, 2007; Chen and Pu 2010a; Chen and Wang 2017; Chen et al. 2019). Moreover, an eye-tracking experiment revealed different eye-gaze patterns on the two interfaces

[1] Technical specifications at https://www.tobiipro.com/product-listing/tobii-pro-x3-120/. The gaze precisions are 0.34 and 0.24 under monocular and binocular conditions, respectively.

(a) LIST-smartphones **(b)** ORG-smartphones

Fig. 1 LIST interface (left) and ORG interface (right) for smartphones (the interfaces for movies and hotels can be found in "Appendix 1")

(Chen and Pu 2010b, 2014): in ORG users viewed more items and more frequently compared products across categories, while in LIST users mainly viewed the items in the top area and rarely attended to items at the bottom of the interface.

Informed by these differences, we deployed the two interfaces in this study, aiming to study their effectiveness for detecting user personality. Also, we implemented each interface for three product domains: *smartphones*, *movies*, and *hotels*. In addition to representing three typical recommendation domains of electronics, entertainment, and tourism (Lu et al. 2015; Ricci et al. 2015), respectively, they reflect three levels of risk associated with the purchase price (Tintarev and Masthoff 2012): smartphones are relatively high risk, movies are low risk, and hotels are normally in-between. According to business and psychology studies (Pachur and Spaar 2015; Nicholson et al. 2005), the effects of personality on individual differences in decision making, e.g., risk-taking propensity and decision style, can be domain-dependent, implying that user behavior can vary across domains. Hence, it is interesting to study how domain differences impact the personality detection process. In summary, we experimented with 6 interfaces (2 designs by 3 domains).

As shown in Fig. 1a, b, the top candidate is the best recommendable item in both the interfaces. Then, in LIST, the remaining items are sorted according to their popularity in a descending order; in ORG, they are grouped into k categories ($k = 4$ in Fig. 1b) as generated by the organization algorithm (Chen et al. 2019). The layout design of ORG

formally adheres to a quadrant structure, where two categories are laid out in parallel, because it was shown that such a layout maximized the difference between the LIST and ORG interfaces with respect to users' eye-movement (Chen and Pu 2010b, 2011, 2014).

We crawled smartphone data from *ZOL.com.cn*[2], a leading IT portal in China. Each product was described by five attributes—price, performance, battery, appearance, and camera—each associated with static product specification and user sentiment[3]. Movie data were crawled from *Mtime.com*[4], a popular movie portal with 170 million unique monthly visitors. Each movie was described by eight attributes: country, release date, type, director, actors/actresses, plot, music, and cinematography. The hotel data were crawled from *Booking.com*[5], a popular accommodation booking website. Each hotel was described by six attributes: price, location, service, facility, WiFi, and cleanliness. An '*Interested*' button was shown next to each product, allowing users to mark the items they liked.

3.2 Personality traits

As mentioned, the Big-Five (Big-5) personality model has widely been used in recommender systems (Tkalcic and Chen 2015), because the five traits of the model are all associated, to a certain extent, with user preferences for items (Rentfrow and Gosling 2003; Cantador et al. 2013; Nguyen et al. 2018; Manolios et al. 2019) as well as their interaction behavior when using RS (Hu and Pu 2013). The Big-5 model is rooted in language analysis, where researchers extracted a set of adjectives describing people's stable traits, further clustered into five factors: Openness to experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (McCrae and John 1992).

- *Openness to experience* (O), referred to as *Openness*, indicates whether a person is creative/open-minded (high O) or reflective/conventional (low O). High O people are typically individualistic, non-conforming, and are aware of their feelings. They have intrinsic propensity to broaden their horizons with new experiences (McCrae and Costa Jr 1997). People with low O tend to have stable interests and prefer simple thinking over complex, ambiguous and subtle.
- *Conscientiousness* (C) inherently leads people to become self-disciplined/prudent (high C) or careless/impulsive (low C). High C people are more likely to be engaged in planned, methodical, hardworking, and achievement-oriented behaviors than low C people (Goldberg 1990).
- *Extroversion* (E) distinguishes sociable/talkative people (high E) from those who are reserved and shy (low E). High E people are usually characterized by poor endurance and resistance, carelessness, and flexibility (Sadi et al. 2011), suggesting

[2] http://mobile.zol.com.cn/.

[3] Explicit user ratings for an attribute were averaged as the sentiment score. If ratings were unavailable, feature-level opinion mining of the product reviews was applied to infer the sentiment score (Chen and Wang 2017).

[4] http://movie.mtime.com/.

[5] http://www.booking.com/.

that they are inclined to rely on intuitive impression for decision making, rather than thinking analytically and logically (Riaz et al. 2012).
- *Agreeableness* (A) reflects individual differences with respect to cooperation and social harmony. People with high A tend to perceive people or things as trustworthy, especially when they encounter positive cues (Goldberg 1990).
- *Neuroticism* (N) reflects an individual's tendency to experience negative feelings. People with high N are less stable emotionally than those with low N. Neuroticism is a predictor of users' tendency to maximize, i.e., seeking the best alternative through systematic and exhaustive searches (Purvis et al. 2011), and high N people tend to feel others having high standards for them and be afraid of receiving negative evaluations (Stoeber et al. 2009).

To measure each participant's personality, we adopted the established 44-item Big-Five Inventory (BFI) (John et al. 1999), due to its strong convergence and discriminant validity. Compared to longer [e.g., NEO PI-R with 240 items (Costa and McCrae 1992) and IPIP with 50/100 items (Goldberg et al. 2006)] and shorter [e.g., TIPI with 10 items (Gosling et al. 2003)] inventories, BFI strikes a balance by avoiding the fatigue effect and ensuring satisfactory psychometric properties (John et al. 1999). Specifically, BFI[6] yields a score for the five personality traits based on 10 *Openness* items, 9 *Conscientiousness* items, 8 *Extroversion* items, 9 *Agreeableness* items, and 8 *Neuroticism* items. Each item is phrased as a short statement "*I see Myself as ..*" rated on a 5-point Likert scale ranging from "strongly disagree" to "strongly agree". For example, the *Openness* items include *"..original, coming up with new ideas"*, *"..curious about many different things"*, *"..ingenious, a deep thinker"*, *"..having an active imagination"*, and more.

3.3 Experimental procedure and participants

We randomly assigned each participant into one out of the above six interfaces (2 interface designs × 3 product domains). The experiment consisted of the following three steps[7], for which an administrator was present during the whole session.

- *Step 1:* The administrator asked the participant to fill out the BFI questionnaire for assessing their personality. Then, the participant was requested to sign a consent form to give the consent to use the eye-tracker to record their eye movements.
- *Step 2:* The administrator introduced the task of evaluating an interface, e.g., "*Imagine you plan to buy a new smartphone, please use the following interface to select two smartphones that you are interested in*." We asked users to choose two items in order to motivate them to compare more options. This is in line with a common shopping behavior, where users often save multiple items to their wish list before making the final choice (Chen 2010).

[6] We adopted a validated Chinese version of BFI in our experiment (Carciofo 2016).

[7] The experiment was originally conducted in a within-subjects design where each participant was asked to interact with both types of interfaces in a random order. For this work, we only considered the first interface they used, in which case the experimental procedure was simplified into three steps.

Table 1 Demographic data of participants

Gender	Male (50), Female (58)
Age	18–25 (98), 26–30 (10)
Pursued education degree	Bachelor (60), Master (43), PhD (5)
Major	Electrical engineering, computer science, industrial design,optics, mechanical engineering, clinical medicine, etc.

– *Step 3:* The participant used the assigned interface to accomplish the task, during which their eye movements were recorded by the eye tracker[8].

We recruited participants through internal email lists and advertisements on social media[9]. One hundred and thirty volunteers participated in the experiment. We filtered out those with calibration difficulties, incomplete eye movement recordings, or users who spent less than 30 seconds using the interface (was deemed too short to make a thoughtful decision Pu and Chen 2007). As a result, we retained 108 users (53.7% females). They were all Chinese students aged 18 to 30, pursuing the Bachelor, Masters or PhD degree at the time of experiment, majors in electrical engineering, computer science, industrial design, mechanical engineering, and more. The summary of the participants' self-reported demographic data is given in Table 1.

Figure 2 illustrates the distribution of the participants' personality trait values. It can be seen that the values are centered between 3 and 5 for *Openness* and *Agreeableness*, between 3 and 4 for *Conscientiousness*, and between 2 and 4 for *Extroversion* and *Neuroticism*. Table 2 provides further descriptive statistics, from which we observe that the mean value for *Agreeableness* is the highest (3.88), and that for *Neuroticism* is the lowest (2.85). The standard deviations are comparable across all the traits. Skewness indicates that all the traits but *Extroversion* are negatively-skewed with a longer tail on the low side of the distribution. Kurtosis shows that *Openness* and *Conscientiousness* have more outliers than in normal distribution. Normality check with the Shapiro–Wilk test reveals that all the traits but *Conscientiousness* are normally distributed ($p > .05$). For the personality predictions, we split the raw values of each trait into the *low* and *high* classes using the median split method (Iacobucci et al. 2015), in order to maintain the balance between the two classes (see the range of values of each class w.r.t. every trait in Table 2). We also analyzed performance of the method for specific interfaces and product domains. As shown in Table 3, the distribution of users into the *low* and *high* classes across different interfaces and domains remains reasonably balanced.

[8] After the calibration procedure, the participants were asked to stay approximately 60–65 cm away from the eye tracker when performing the task, as per the eye-tracker's manual.

[9] The experimental procedure was approved by the University Research Ethics Committee.

Table 2 Statistical analysis of participants' personality trait values

Trait	Descriptive statistics				Normality test			Two classes	
	Mean	Std	Skewness	Kurtosis	Stat.	*df*	*p*	Low	High
Openness	3.59	.56	−.440	.550	.979	108	.078	[1.80, 3.60]	[3.70, 5.00]
Conscientiousness	3.32	.54	−.716	1.659	.960	108	.002	[1.33, 3.22]	[3.33, 4.56]
Extroversion	3.10	.59	.341	−.220	.978	108	.069	[1.75, 3.00]	[3.13, 4.75]
Agreeableness	3.88	.53	−.369	−.078	.984	108	.225	[2.22, 3.89]	[4.00, 5.00]
Neuroticism	2.85	.67	−.184	−.158	.991	108	.707	[1.13, 2.75]	[2.88, 4.50]

Table 3 Numbers of users in the *low* and *high* classes of each experimental condition

	ALL		Recommendation interface				Product domain					
			LIST		ORG		Smartphones		Movies		Hotels	
	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High
Openness	51	57	23	29	28	28	19	17	14	22	18	18
Conscientiousness	51	57	24	28	27	29	19	17	16	20	16	20
Extroversion	55	53	26	26	29	27	19	17	19	17	17	19
Agreeableness	53	55	23	29	30	26	15	21	20	16	18	18
Neuroticism	50	58	24	28	26	30	15	21	17	19	18	18

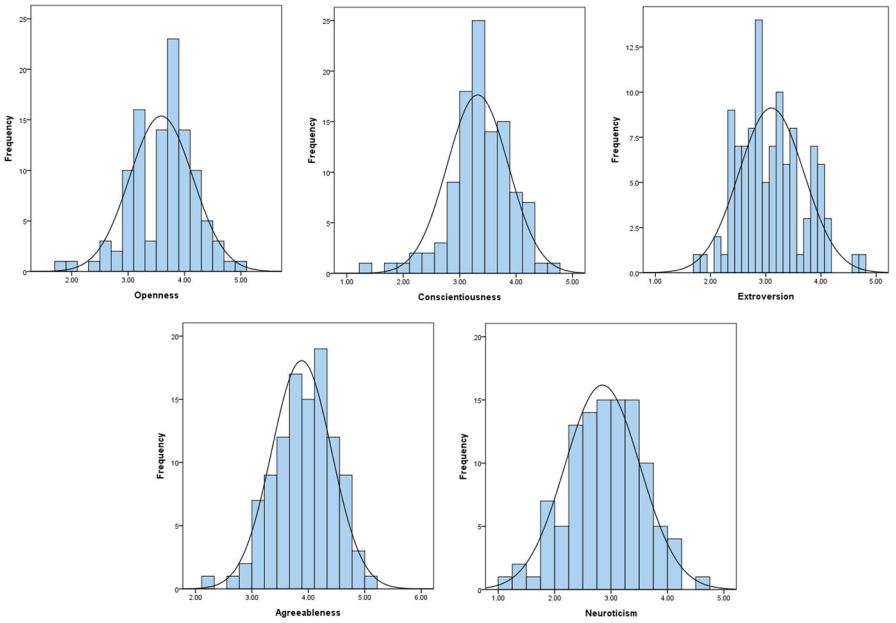


Fig. 2 Distribution of personality trait values among our participants

3.4 Eye-movement features

We extracted features pertaining to the users' eye movements over the entire interface as well as being related to more fine-grained Areas of Interest (AOIs) (Poole and Ball 2005). Specifically, an AOI is defined at two levels:

- *Group level:* Because in ORG all the products except the top candidate are grouped into four categories, we naturally divided the interface into five group-level AOIs, as shown in Fig. 3. Specifically, the region that displays the top candidate is considered as an individual AOI, and each category corresponds to an AOI accommodating its six contained products. To obtain similar group-level AOIs in LIST, we divided it into five AOIs: the top candidate, and four groups of six products each, i.e., ranked 2 to 7, 8 to 13, 14 to 19, and 20 to 25.
- *Product level:* At a more fine-grained level, each recommended product represents a separate AOI, allowing to obtain eye-movement data at the product level (see Fig. 4). In this case, each AOI covers the area associated with one particular product, including its image and textual description.

Although the definition of the AOIs was handcrafted in this experiment, it is important to stress that the process can be automated in real applications. For example, web page panels and specific interface elements can be extracted from the layout of a HTML page. In this way, the system can automatically identify the locations of the group-level and product-level AOIs, as well as other salient components of the interface.

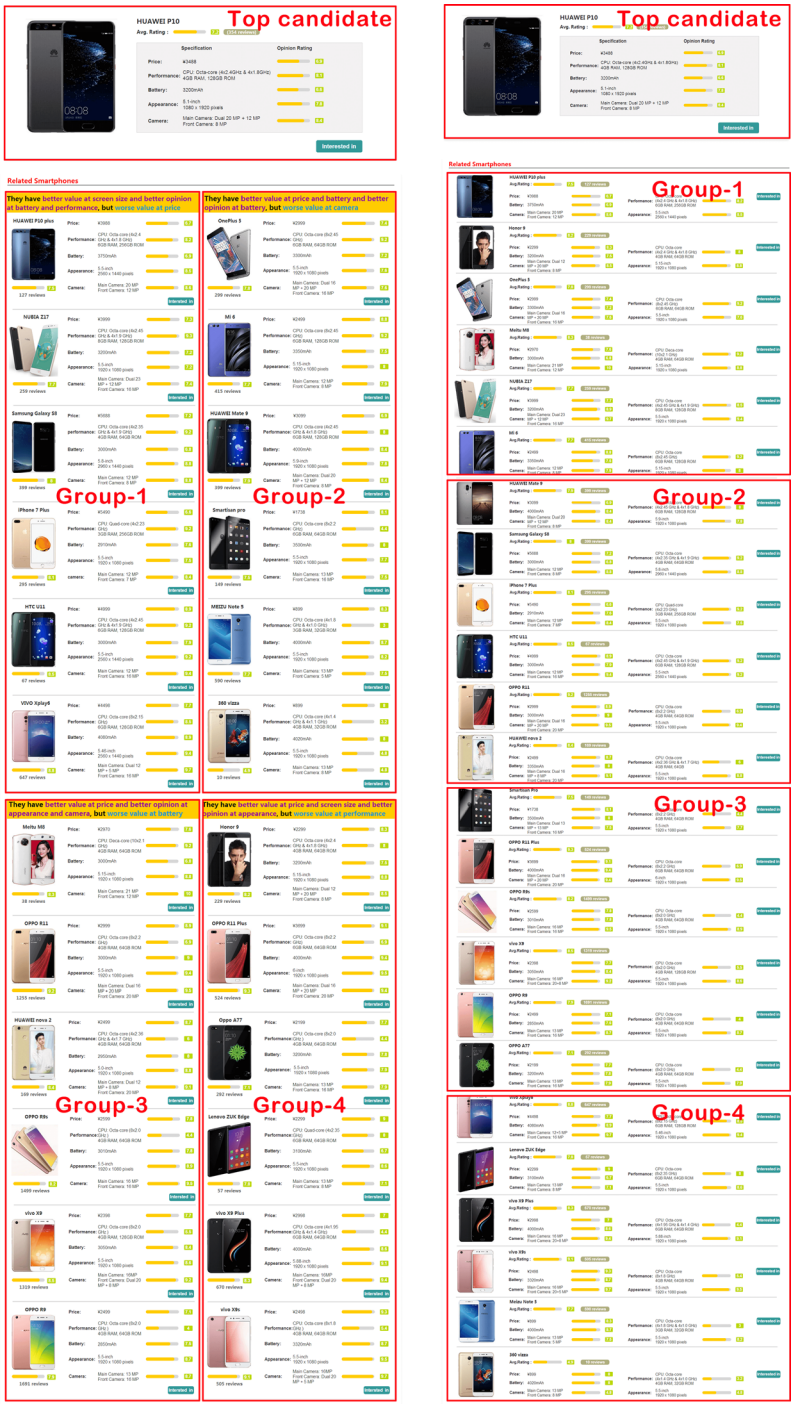


Fig. 3 Group-level AOIs in ORG (left) and LIST (right). In ORG each category represents an AOI, and in LIST all the products (except the top candidate) are evenly divided into four group-level AOIs

Fig. 4 Product-level AOIs in ORG (left) and LIST (right). Each AOI corresponds to one product

We collected the participants' raw eye-movement data through the eye-tracker[10]. The data consist of three types of events: *fixation*— stationary gaze position, with a minimum duration of 200 ms (Salvucci and Goldberg 2000); *saccade*—rapid simultaneous movement of both eyes; and *pupillary response*—changes in pupil sizes (Cavanagh 2014; Kret 2019). We extracted a total of 86 eye-movement features and divided them into three groups: *whole interface*, *group-level*, and *product-level* (see Table 4). For the whole interface, we mainly considered the overall fixation count/duration, fixation rate, dwell time, average fixation duration, and time to the first fixation. Features related to saccade and pupil size were also populated. At the group level, we highlight AOI-specific fixation features, such as the proportion of fixation count/duration on the target AOI relative to all the AOIs. Features pertaining to back-and-forth transitions between AOIs, e.g., group-X to group-Y and then back to group-X, are emphasized, as they may disclose comparisons of products across the AOIs (Chen and Pu 2011). Due to the high number of products, at the product level, we mainly populated the number of unique products a user has fixated on, average fixation count/duration per product, and the total number of back-and-forth transitions among products. We also counted fixations on the two products, respectively, selected by the user, as these selected products naturally attracted more attention and the associated features might be informative.

These features are in line with the eye-tracking metrics considered in related work on understanding users' decision behavior and cognitive load. For instance, fixation measures (e.g., number/duration of fixations, visit count, transitions) and saccade measures (e.g., saccade-fixation ratio, saccade amplitude) have often been used to uncover information search and uptake processes in reading (Toker et al. 2019) and decision making (Franco-Watkins and Johnson 2011; Rojas et al. 2020; Ashby et al. 2016). Pupillary responses (e.g., pupil size and dilation) were shown to be valid indicator of cognitive load within a task (Cavanagh 2014; Kret 2019). Notably, saccade and fixation features were found to be related to user personality in Berkovsky et al. (2019). Thus, we set out to investigate how a broader range of eye-tracking features as populated from a user's interaction with a typical recommendation interface might contribute to the detection of their personality traits.

3.5 Personality prediction

The predictions of users' personality traits from their eye movements can be split into two steps: *feature selection* and *classification*. As listed in Table 4, we used the captured eye-gaze interaction with a recommendation interface to populate 86 features. In order to avoid overfitting, we used feature selection that picked a subset of most predictive features in training data. Specifically, we deployed five feature selection methods (Li et al. 2017): *Gini-Index* score (GI)—assesses if a feature can separate users between the target classes; *Correlation-based feature selection* (CFS)—identifies features correlated with the predicted class label and not correlated with other features; *F-score* (FS)—selects features correlated with the predicted class label

[10] The Tobii I-VT fixation filter was used. During the filtering process, if there were no gaze data within two consecutive seconds in a recording, this recording was removed.

Table 4 List of eye-movement features for personality prediction

Level	Feature	No.	Description
Interface	Fixation count/duration	1–2	Total count/duration of all the fixations on the whole interface
	Fixation rate	3	Average number of fixations per second
	Dwell time	4	Overall gaze time out of the whole interface display time
	Average fixation duration	5	Average duration of fixations
	Time to first fixation	6	Period of time before the first fixation
	Saccade rate	7	Average number of saccades per second
	Saccade amplitude	8	Average angular distance of all the saccades in the interface
	Saccade-fixation ratio	9	Ratio between the duration of saccades and that of fixations
	Average pupil size	10	Average diameter (in pixels) of the left and right pupils
	Pupil dilation	11	Change in pupil size between the first and last fixations on the interface
Group	Visit count	12–16	Number of visits to the AOI (a visit indicates that a user fixates on the AOI from outside)
	Fixation count	17–21	Total number of fixations on the AOI
	Proportion of fixations	22–26	Fixations on the AOI out of all the fixations in all the AOIs
	Fixation rate	27–31	Average number of fixations per second on the AOI
	Dwell time	32–36	Overall gaze time for the AOI out of the whole interface display time
	Fixation duration	37–41	Total duration of fixations on the AOI
	Proportion of fixation duration	42–46	Duration of fixations on the AOI out of the total fixation duration
	Average fixation duration	47–51	Average duration of fixations on the AOI
	Time to the first fixation	52–56	Period of time before the first fixation on the AOI
	Longest fixation	57–61	Longest fixation duration on the AOI
	Back-and-forth transition count	62	Total number of back-and-forth transitions across all the AOIs
	Back-and-forth group transition count	63	Total number of back-and-forth transitions across the four group-level AOIs
	Proportion of group transitions	64	Back-and-forth group transitions out of all back-and-forth transitions
	Back-and-forth top-group transition count	65	Total number of back-and-forth transitions from the top candidate to the four group-level AOIs

Table 4 continued

Level	Feature	No.	Description
	Proportion of top-group transitions	66	Back-and-forth top-group transitions out of all back-and-forth transitions
	Back-and-forth group-top transition count	67	Total number of back-and-forth transitions from the four group-level AOIs to the top candidate
	Proportion of group-top transitions	68	Back-and-forth group-top transitions out of all back-and-forth transitions
Product	Fixated products count	69	Total number of unique products that were fixated on
	Average fixation count/duration	70–71	Average count/duration of product fixations
	Back-and-forth transition count	72	Total number of back-and-forth transitions across all the products
	Fixation count on the 1st/2nd choice	73–74	Total number of fixations on the selected product
	Fixation duration on the 1st/2nd choice	75–76	Total duration of fixations on the selected product
	Fixation rate on the 1st/2nd choice	77–78	Average number of fixations per second on the selected product
	Average fixation duration on the 1st/2nd choice	79–80	Average duration of fixations on the selected product
	Decision time on the 1st/2nd choice	81–82	Period of time before choosing the product
	Average fixation count/duration on unselected products	83–84	Average count/duration of fixations on the non-selected products
	Max. fixation count/duration on unselected products	85–86	Maximal count/duration of fixations on the non-selected products

by calculating ANOVA F-value; *T-score* (TS)—measures if a feature can make the mean values between the target classes statistically different; and *Fisher score* (FIS)—selects features, for which users within a class are similar and users across classes are dissimilar.

Owing to the median split, the *low* and *high* classes were balanced, allowing us to deploy binary classifiers and predict the class labels. We experimented with nine classifiers, including AdaBoost (AB), XGBoost (XGB), Gradient Boosting Decision Tree (GBDT), Random Forest (RF), Decision Tree (DT), Naive Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP), all of which were implemented with the `scikit-learn`[11] and `scikit-feature`[12] Python packages (the code of our implemented methods is released[13].). For each classifier, we first used fivefold cross-validation to tune the

[11] https://scikit-learn.org.

[12] https://github.com/jundongl/scikit-feature.

[13] https://github.com/wanlingcai1997/personality_prediction_code.git.

hyper-parameters on the 80% training data as randomly sampled from the whole data, where the *Accuracy*[14] metric was used to identify the optimal values of the hyper-parameters. We then conducted tenfold cross-validation to evaluate the models with those identified hyper-parameters. The reported Accuracy score is the average over all users in the test set.

4 Results

In this section, we report the prediction results, which are broken into two major analyses. *First*, we identify the most informative features for the prediction of every trait and the best-performing classification algorithms. *Second*, we study interface and domain dependencies, i.e., how the performance changes for the LIST and ORG interfaces, as well as for smartphones, movies, and hotels.

4.1 Feature selection and classification

To evaluate the feature selection methods, we first fused the five trait-specific classifiers into one metric by averaging the prediction accuracy over the five traits. We increased the number of selected features from 2 to 40 with a step of 2 for GI, FS, TS, and FIS[15]. Because CFS automatically selects the optimal feature subset, this parameterization was not necessary. We also compared with the baseline setting using all the 86 features without feature selection.

The obtained accuracy scores are listed in Table 5. The results show that the accuracy of all the methods but CFS is superior to that of the classifier using all the available features. This clearly indicates that selecting the informative features is beneficial for the personality predictions and improves their accuracy. Comparing the GI, FS, TS, and FIS feature selectors, we observe that GI achieves the highest accuracy for six classifiers: AB, XGB, GBDT, RF, DT, and MLP. For the remaining classifiers, TS achieves the highest accuracy (for NB, on par with FS and FIS).

We conducted another experiment, where the number of features selected by each classifier was not fixed, but rather optimized by the classifier. For the GI, FS, TS, and FIS feature selectors, the performance of the five trait-specific classifiers was again fused into a single metric reported in Table 6. CFS is not shown, as the number of features was not fixed in Table 5, and the results would be identical. Consistently with Table 5, the results show that, when the number of features is not fixed, GI outperforms the other methods for all the classifiers but NB and LR. Comparing the last average rows in Tables 5 and 6, we note that the accuracy degrades when the number of features is fixed, suggesting that the optimal number of features varies across the traits. We

[14] Accuracy refers to the proportion of correct predictions (i.e., *low* or *high* class label being predicted) among all the predictions.

[15] In another experiment, we varied the number of features from 10 to 80 with a step of 10. The results showed that the highest accuracy was achieved when the number of features is below 40.

Table 5 Comparison of feature selection methods

Classifier	GI	CFS	FS	TS	FIS	All features
AB	**0.6773 (8)**	0.5511	0.6544 (26)	0.6506 (32)	0.6544 (26)	0.6229
XGB	**0.6331 (12)**	0.5483	0.6031 (22)	0.5905 (10)	0.6031 (22)	0.5409
GBDT	**0.6400 (28)**	0.5230	0.6125 (18)	0.6077 (10)	0.6037 (22)	0.5520
RF	**0.6541 (4)**	0.5302	0.6205 (10)	0.6168 (26)	0.6163 (26)	0.5547
DT	**0.6562 (4)**	0.5055	0.5866 (26)	0.5833 (8)	0.5822 (24)	0.5546
NB	0.5735 (2)	0.4920	**0.5878 (14)**	**0.5878 (14)**	**0.5878 (14)**	0.4959
LR	0.6171 (30)	0.5194	0.6313 (16)	**0.6402 (16)**	0.6323 (16)	0.5471
SVM	0.6116(8)	0.4952	0.6048 (16)	**0.6140 (14)**	0.6079 (18)	0.5218
MLP	**0.6143 (20)**	0.5471	0.5945 (2)	0.5982 (22)	0.5868 (16)	0.5810
Average	**0.6308**	0.5235	0.6106	0.6099	0.6083	0.5523

The number of features (fixed for all traits) yielding the best performance is given in the bracket. For every classifier, the best performing method is in bold
bold highlights the highest value in each row

Table 6 Comparison of feature selection methods

Classifier	GI	FS	TS	FIS
AB	**0.7043**	0.6817	0.6782	0.6817
XGB	**0.6745**	0.6361	0.6306	0.6361
GBDT	**0.6825**	0.6516	0.6498	0.6461
RF	**0.6803**	0.6402	0.6494	0.6429
DT	**0.6831**	0.6432	0.6592	0.6406
NB	0.6027	**0.6114**	0.6045	**0.6114**
LR	0.6525	**0.6613**	0.6580	0.6591
SVM	**0.6533**	0.6331	0.6286	0.6322
MLP	**0.6570**	0.6241	0.6422	0.6373
Average	**0.6656**	0.6425	0.6445	0.6430

The number of features is not fixed across the traits
bold highlights the highest value in each row

posit that the superiority of GI in both cases can be explained by its measurement of impurity[16] for recognizing distinguishable features that can accurately classify users.

Next, we turn to the classification methods deployed for the predictions of personality traits. Table 7 shows the results of the nine classifiers with the GI feature selector for predictions of the five traits. We observe that AB achieved the best accuracy for *Openness*, *Extroversion*, and *Agreeableness* (for the first two, on par with GBDT or DT). For *Conscientiousness* and *Neuroticism*, AB was outperformed by RF and LR, respectively. Considering individual traits, we observe that all traits but *Extroversion* were predicted with accuracy greater than 0.7. Averaging performance across all the classifiers, *Conscientiousness* yielded the highest accuracy. This can be attributed to

[16] Impurity measures how often a random element is incorrectly labeled according to the class distribution in the data.

Table 7 Comparison of nine classifiers for the five personality traits

Classifier	Openness	Conscientiousness	Extroversion	Agreeableness	Neuroticism	*Average*
AB	**0.7041**	0.7329	**0.6682**	**0.7197**	0.6964	**0.7043**
XGB	0.6774	0.7221	0.5959	0.6905	0.6864	0.6745
GBDT	**0.7041**	0.7221	0.6491	0.6588	0.6782	0.6825
RF	0.6933	**0.7420**	0.6165	0.6523	0.6973	0.6803
DT	0.6865	0.7412	**0.6682**	0.633	0.6864	0.6831
NB	0.542	0.6558	0.5423	0.6615	0.6118	0.6027
LR	0.6145	0.6674	0.5956	0.6633	**0.7218**	0.6525
SVM	0.6412	0.6758	0.6306	0.6524	0.6664	0.6533
MLP	0.6103	0.6871	0.6223	0.6508	0.7145	0.6570
Average	0.6526	**0.7052**	0.6210	0.6647	0.6844	

bold highlights the highest value in each column

the analytical nature of our decision making task, which required the participants to compare products. We posit that *Conscientiousness* is the trait, where analytical skills manifest most in the Big-5 model (Poropat 2009).

Across the five traits, the average accuracy of AB was 3–3.5% higher than DT, GBDT, and RF, and substantially higher than the accuracy of other classifiers. We ran the one-way repeated-measures ANOVA test[17] on the averages of these methods for all traits, which shows that there is a significant difference among the nine classifiers ($F = 5.170$, $p < 0.001$). The post hoc pairwise dependent sample t test further shows that the two classifiers AB and NB are significantly different ($t = 3.870$, $p = 0.012$ after Bonferroni correction[18]).

There are two commonalities among the four classifiers: AB, DT, GBDT, and RF. *First*, their best accuracy in Tables 5 and 6 was achieved with the GI feature selection method. *Second*, they are all tree-based classifiers, implying that tree-based classifiers might complement GI's feature selection process and excel in discovering nonlinear relationships among the features. Overall, we conclude that the ***AB classifier combined with GI feature selector*** predicts the traits more accurately than other combinations, and we will focus on this combination in the subsequent experiments.

4.2 Recommendation interface and application domain

In the next analysis, we set out to identify what recommendation interface—ORG or LIST—is more informative for personality trait predictions. For this, we divided the entire data into two groups: users who used the LIST interface and those who used

[17] This test was chosen owing to its ability to determine whether three or more group means (i.e., the nine classifiers in our case) are significantly different, where the participants are the same in each group (Howell 2012). We further conducted post hoc dependent sample t test for pairwise comparisons. All the reported significance tests were performed on the tenfold cross-validation results.

[18] As there were a total of 36 pairwise comparisons among the 9 classifiers, the Bonferroni-corrected p value was calculated by multiplying the uncorrected p value by 36.

Table 8 Comparison of the ORG and LIST interfaces for predictions of the five personality traits

Personality trait	ORG	LIST	Combined
Openness	0.7000	**0.7900**	0.7041
Conscientiousness	0.7050	**0.7833**	0.7329
Extroversion	**0.6967**	0.6750	0.6682
Agreeableness	**0.7400**	0.6733	0.7197
Neuroticism	0.7333	**0.8200**	0.6964
Average	0.7150	**0.7483**	0.7043

AB (classifier) and GI (feature selector) are the deployed methods
bold highlights the highest value in each row

ORG. We then used the features associated with one interface only to train the classifier and predict personality traits. Given our previous conclusions about the performance of the AB classifier combined with GI feature selector, we report the performance of this combination only.

As shown in Table 8, LIST facilitated more accurate predictions than ORG, with the respective average accuracy values of 0.7483 and 0.7150 across all the traits. Namely, LIST outperformed ORG for *Openness*, *Conscientiousness*, and *Neuroticism*, while ORG achieved higher accuracy for *Extroversion* and *Agreeableness*. On average, the predictions produced using LIST were more accurate by 4.65%, which might be explained by more fixation data captured by this interface (totally 7577 gaze data points and 145.71 per user vs. 6,682 and 132.03 per user in ORG) allowing to extract more informative features. We attribute this to the less organized nature of LIST that naturally demanded more interactions. Overall, ORG and LIST exhibit different strengths in predicting different traits, which might require further investigation as the differences between the two interfaces are not significant according to independent samples t test[19] ($p > 0.05$ across all the traits).

We also compared the predictions using either the ORG or LIST interface to the combined group reported in Table 7. It is evident that the best interface-specific predictions were consistently more accurate than the combined ones. Moreover, for *Extroversion* and *Neuroticism*, even the predictions with the lower-performing interface were superior to the combined ones. Hence, the average accuracy values of both the LIST and ORG interfaces (respectively, 0.7483 and 0.7150) were higher than the combined average accuracy (0.7043). Note that the interface-specific accuracy for most traits is above 0.7, which is comparable with the results of Big-5 personality predictions reported in prior literature (Majumder et al. 2017; Wache et al. 2015). For *Openness*, *Conscientiousness* and *Neuroticism*, the accuracy is close to 0.8, as these were shown to relate to users' decision behaviors (Pachur and Spaar 2015; Purvis et al. 2011). This implies that the accuracy of personality predictions depends on the recommendation interface design.

In order to identify what groups of features and AOIs play an important role in the prediction process, we retrieved all the selected features for each trait, together with their relative importance, i.e., weight, determined by AdaBoost (AB). For this, we

[19] We chose this test to compare the means of two independent groups (two recommendation interfaces in our case), as the participants are different between the two groups.

added up the weights of features at the whole interface, group level, and product level, respectively, as defined in Table 4. Figure 5a, c shows the aggregated importance of the selected features at each level. It can be seen that group-level features dominate those extracted from the whole interface and individual products, with their aggregated importance ranging from 0.40 to 0.64 in ORG and from 0.42 to 1 in LIST. Second most-predictive features are at the product level, ranging from 0.34 to 0.46 for predictions of four traits in ORG and having approximately half of the weight for two traits in LIST. Notably, interface-level features have the lowest weight and their aggregated importance does not pass 0.3 across all the traits and interfaces.

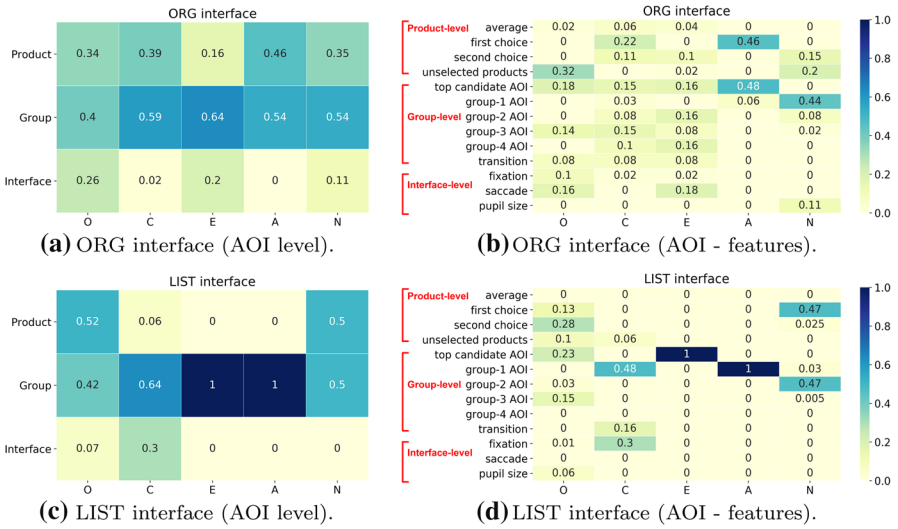
Further breaking the features into different levels of AOIs associated with them and summing up the weights of the selected features, we note in Fig. 5b, d two interesting findings. *First*, the dominant features are mainly associated with the first-choice item at the product level, or the top item, group-1, or group-2 AOIs at the group level. This shows that the most informative features can be extracted from user interactions with products displayed in the top area, consistently with previous findings showing that users pay more attention to the top area (Chen and Pu 2010b). In this study, we observed 5879 gaze data points (113.06 per user) in the top area including the top item and the first two groups, while only 1698 (32.65 per user) in the bottom area of LIST. Likewise, in ORG there were 5126 data points (91.54 per user) in the top area and 1556 (27.79 per user) in the bottom. *Second*, for both interfaces, we note the strong dominance of these features for *Agreeableness* and *Neuroticism*, suggesting that these traits are less sensitive to the interface differences. For *Openness*, although the feature weight distributions are similar across the interfaces, the weights are more balanced. Recall that *Openness* inherently determines a person's propensity to explore new options, which could explain the higher number of interactions across various AOI levels.

We were also interested to investigate whether the product domain would affect personality predictions. For this, the whole dataset was divided according to the domain, with which a subject interacted, i.e., smartphones, movies, or hotels. Table 9 shows the results obtained for each domain, still using GI for feature selection and AB for classification. It can be observed that the highest average accuracy was obtained for Hotels (0.7900), followed by Movies (0.7653), and then Smartphones (0.6923). Considering the traits individually, we note that *Openness* and *Conscientiousness* were

Table 9 Personality trait prediction in different product domains

Personality trait	Smartphones	Movies	Hotels	Combined
Openness	**0.7333**	0.7183	0.7000	0.7041
Conscientiousness	**0.8417**	0.7167	0.8083	0.7329
Extroversion	0.6167	0.6167	**0.7667**	0.6682
Agreeableness	0.7017	**0.9417**	0.9000	0.7197
Neuroticism	0.5683	**0.8333**	0.7750	0.6964
Average	0.6923	0.7653	**0.7900**	0.7043

AB (classifier) and GI (feature selector) are the deployed methods
bold highlights the highest value in each row



(a) ORG interface (AOI level). **(b)** ORG interface (AOI - features).

(c) LIST interface (AOI level). **(d)** LIST interface (AOI - features).

Fig. 5 Relative importance of the selected features across the two interfaces

predicted most accurately in the Smartphone domain (0.7333 and 0.8417, respectively), *Agreeableness* and *Neuroticism* were predicted best in the Movie domain (0.9417 and 0.8333, respectively), and the most accurate prediction of *Extroversion* was achieved in the Hotel domain (0.7667). We conducted a one-way ANOVA test for each personality trait to compare the results across the three product domains[20], which shows significant differences with respect to two traits: *Agreeableness* ($F = 4.860$, $p = 0.015$) and *Neuroticism* ($F = 5.170$, $p = 0.012$). The post hoc Tukey's HSD test additionally shows that the differences between Smartphones and Movies are significant in terms of predicting *Agreeableness* ($t = 2.917$, $p = 0.011$) and *Neuroticism* ($t = 3.060$, $p = 0.007$), and those between Smartphones and Hotels are also significant for predicting *Agreeableness* ($t = 2.411$, $p = 0.045$) and *Neuroticism* ($t = 2.386$, $p = 0.048$).

It is important to emphasize that for three traits—*Conscientiousness*, *Agreeableness*, and *Neuroticism*—the predictive accuracy surpassed 0.80 when the product domain was taken into consideration. This result is comparable, and even superior, to the results reported in other works focusing on Big-5 predictions (Majumder et al. 2017; Wache et al. 2015). In particular, the highest accuracy was obtained for *Agreeableness* prediction in the Movie domain, reaching as high as 0.9417. Comparing domain-specific predictions to the combined ones reported in Table 7, we observe that the best domain-specific predictions outperformed the combined ones, consistently across all the traits. Moreover, for all the traits but *Extroversion*, the second-best domain-specific predictions were still superior to the combined predictions. In agreement with this, the average accuracy values obtained for the Hotel and Movie domains (respectively, 0.7900 and 0.7653) were higher than the 0.7043 average accuracy of the

[20] The one-way ANOVA test was used for comparing more than two independent groups (i.e., three product domains in our case) (Howell 2012).

combined predictions. These results imply that the accuracy of personality predictions depends also on the application domain.

In similar to the analysis of features across the interfaces, we retrieved all the selected features as well as their weights for each trait in domain-specific predictions. As shown in Figs. 6a, c, e, group-level and product-level features generally performed better than interface-level features. Further analyses considering individual feature types, as shown in Fig. 6b, d, f, indicate that for *Conscientiousness*, the features were more evenly distributed in the Smartphone domain, while *Openness* exhibits such pattern mainly in the Movie domain, which shows the sensitivity of these two traits to domain properties. We posit that this observation might be (partially) explained by the risk level associated with the domain, as implied by related decision-making studies in psychology (Pachur and Spaar 2015). Specifically, for high-risk products such as smartphones, high *Conscientiousness* people might be inclined to use deliberative, i.e., effortful, planned, and analytic, decision mode, whereas for low-risk products, e.g., movies, high *Openness* people might be driven to explore diverse and new items. Conversely, *Extroversion* and *Neuroticism* are relatively domain-insensitive, as many selected features were evidently associated with the selected items, the top candidate, and the top groups. It is also interesting to note domain differences in the predictions of *Agreeableness*, which will require a deeper investigation.

5 Discussion

In recent years, more attention has been paid to building *personality-based recommender systems* (RS) because it was found that personality can inherently affect user preferences and interaction behavior (Rentfrow and Gosling 2003; Hu and Pu 2013; Cantador et al. 2013; Manolios et al. 2019). However, most of the existing studies acquire users' personality via psychometric questionnaires, which unavoidably put burden on users and potentially raise reliability concerns, being prone to manipulation due to their self-reported nature (Anglim et al. 2018; Fahey 2018). Hence, in this work, we have investigated the feasibility of implicitly and objectively acquiring users' personality traits from their eye movements over a recommendation interface, given that eye-tracking is deemed to be a useful process-tracking tool for capturing users' cognitive processes associated with decision making (Franco-Watkins and Johnson 2011; Glaholt and Reingold 2011; Ashby et al. 2016).

In this work, we experimentally tested the performance of 9 classifiers, applied in conjunction with 5 feature selectors, for predictions of Big-5 personality traits. Overall, we collected a rich dataset of more than 14,000 gaze data points from 108 subjects, as captured from their first interaction with a recommendation interface that was randomly associated with one product domain among three. The obtained results show that the AdaBoost (AB) classifier combined with the Gini-Index score (GI)-based feature selector predicted the personality traits more accurately than other combinations. In particular, the *Openness*, *Conscientiousness*, and *Agreeableness* traits were predicted better than the rest, in line with related work linking them to cognition and decision behaviors (Zillig et al. 2002).

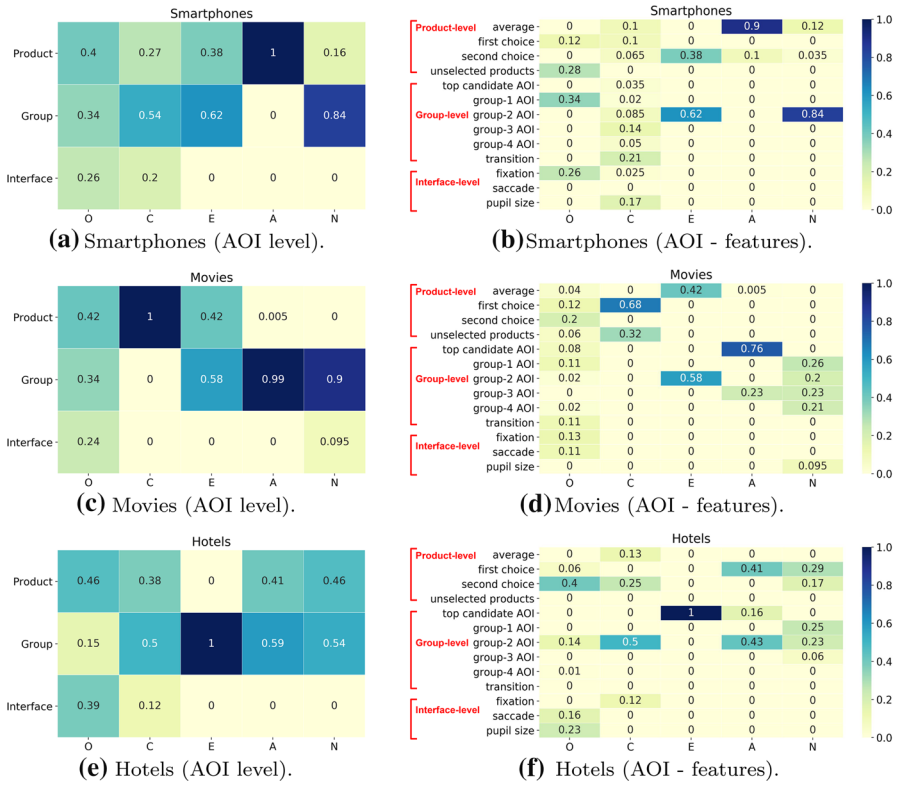


Fig. 6 Relative importance of the selected features across the three domains

We also restricted the data collection to specific interfaces or to interaction with specific application domains. Such interface- and domain-specific predictions achieved a higher accuracy than the combined ones, although we could not decisively conclude on the most predictive interface or domain. In other words, certain interfaces and domains were informative for predicting certain personality traits and failed to capture information allowing to accurately predict other traits. This brings to the fore an important question regarding the links between interfaces/domains and personality traits. For example, the structured nature of the ORG interface may be more appealing for highly conscientious users, or alternatively, the artistic nature of movie thumbnails may attract the attention of high-openness users. This may be evident in their eye-tracking data and help the classifier to accurately predict the values of these traits. However, such examples are likely to be hand-picked and, at this stage, our results with respect to interface and domain dependencies preclude us from drawing strong conclusions.

Another intriguing question refers to the volume of interactions required to reliably detect personality traits. Recall that our personality detection method was conceived as a means to model new users and bootstrap personalization, while rich user data are not available yet. Ideally, we would like to shorten the bootstrapping and provide per-

sonalized recommendations as early as possible (Golbandi et al. 2010), which brings to the fore the characterization of the bootstrapping interactions. For example, how long is the minimal interaction required to reliably detect personality? How diverse should such an interaction be, in terms of both eye activity and interface components? How is the detection of various traits affected by the bootstrapping interaction? We leave these questions beyond the scope of our work and intend to address them in subsequent analyses.

The obtained results are comparable to (and, in certain cases, outperform) those reported in prior literature (Majumder et al. 2017; Azucar et al. 2018; Wache et al. 2015; Li et al. 2014; Berkovsky et al. 2019; Hoppe et al. 2018). In particular, Li et al. (2014) used digital records of micro-blogging behaviors to predict personality, which achieves classification accuracy slightly higher than ours, i.e., ranging from 0.84 to 0.92 for the Big-5 personality traits (each classified as *high* or *low*). Majumder et al. (2017) relied on stream-of-consciousness textual essays to predict users' personality (each trait classified as *positive* or *negative*), the accuracy of which is the highest (0.63) for predicting *Openness*, while those for the other traits are all below 0.60.

It is worth highlighting that, in related work with the eye-tracker (Wache et al. 2015; Taib et al. 2020), the personality traits were mostly predicted based on responses to affective images and video clips. On the contrary, in this work we did not deploy affective stimuli, but rather relied on observable interactions with typical RS interfaces. We deem this to be both a strength and a limitation. On the one hand, general physiological signals, which are not triggered by affective stimuli, may not have the validity of autonomic nervous system responses. This may increase the risk of overfitting that we tried to mitigate by applying feature selection and verifying the results with multiple classifiers. On the other hand, in our work the signals were captured as part of a natural user interaction with a typical interface deployed by numerous RS in a range of domains, which substantially simplifies the implementation of the work in a practical setting (Hoppe et al. 2018).

Another important limitation refers to the psychological interpretation of the obtained results. As noted earlier, our subject set was relatively homogeneous: Chinese students aged from 18 to 30, mainly having a science or engineering background, and with a close to normal distribution of personality trait values (see Fig. 2). Thus, our cohort included very few subjects positioned at the high or low ends of the personality scale. In addition, the data-driven median split of subjects into two personality classes did not allow us to single out subjects with extreme trait values and potential personality disorders (Morey et al. 2000). Due to the homogeneous background of our subjects and the scarcity of extreme personality trait values in our data, it remains unclear whether our findings regarding the feasibility of eye-tracking-based personality prediction will be valid for different populations and whether the predictions will identify clinical psychology cases or subjects with personality disorders.

From the computational angle, we observed substantial accuracy differences across personality traits, recommendation interfaces, and application domains. Not only the performance of the classifier varied, but also the selected predictive features and their weights fluctuated, as shown in Figs. 5 and 6. We attribute this observation to the relatively small sample size, especially considering the split of the subjects into two interface-specific and three domain-specific datasets. Also, the observed variations

potentially reflect the impact of the participants' familiarity with the recommended items. For example, if a subject is familiar with a recommended movie, this may attract their attention and bias their interaction with the interface (Lancry-Dayan et al. 2018). For obvious reasons, our recruitment could not control for such biases and neither we could eliminate them in the analysis.

It should also be highlighted that a single eye-tracking technology was deployed. This differentiates us from related works that, in addition to an eye-tracker, used skin conductance sensors, electroencephalogram (EEG) and electrocardiogram (ECG) devices, and face trackers (Hoppe et al. 2018; Wache et al. 2015; Tai et al. 2006; Sharan et al. 2020). Despite the limited physiological data captured by the eye-tracker, our results were comparable and in some cases even surpassed those reported in previous works using an array of sensing technologies. This shows the strong potential of our approach, which, if enriched by other sensing technologies, may achieve a substantially higher predictive accuracy. In addition, we note that in the last couple of years several works on simple and affordable eye-tracking technologies, e.g., using web cameras, have been published (Wang and Ji 2017; Bott et al. 2017; Mounica et al. 2019). We posit that the improving accuracy of such technologies may soon eliminate the need for a dedicated sensing device and further streamline the adoption of such technologies.

Last but not the least, we highlight the ethical considerations associated with our work. The developed method facilitates accurate and objective modeling of users' personality, which can then be harnessed to provide better recommendations to users. At the same time, it offers a powerful tool that, if misused, may entrench biases and potentially discriminate (McClendon et al. 2019). For example, *Openness* and *Neuroticism* were shown to correlate with impulsive buying behavior (Shahjehan et al. 2012), such that the knowledge of personality traits can be misused to promote sales. Add to this the possibility of covert collection of eye-tracking data with common technologies, such as web cameras (Bott et al. 2017), and the practical risk of such a technology becomes evident. Technology developers need to be aware of such risks and implement transparent and explainable services to mitigate them and offer some degree of protection to users (Van Nuenen et al. 2020; Kim et al. 2020).

Despite the above, we believe that our findings surface important insights for enhancing personality-based RS. We posit that the personality traits detected in the users' initial interaction with a RS could be harnessed in at least three ways: (1) assist collaborative RS to generate recommendations for new users, e.g., by means of calculating personality-based similarity among them (Tkalcic et al. 2009); (2) enhance recommendation diversity, as an extension of our previous work based on explicit personality acquisition (Wu et al. 2018); and (3) develop cross-domain recommendations (Fernández-Tobías et al. 2016), so that the personality learnt from users' behavior in one domain (e.g., movies) could be exploited to generate personality-aware recommendations in another domain (e.g., smartphones).

Although current research on eye-gaze-based interaction has primarily been conducted in the laboratory, the increased sophistication, accessibility, and accuracy of eye-tracking technologies may upgrade them into a new type of commonly used input devices compatible with computers and smartphones (Zhang et al. 2017; Valtakari et al. 2020). The idea of leveraging eye movements to predict user personality could, hence, be applicable to commercial RS in the future. For instance, imagine a real-time

eye-tracking-based personality detector being integrated into a product RS. This way, not only the recommendation algorithm can be adjusted by considering the detected personality traits, but also the interface can be customized to better meet users' preferences (Alves et al. 2020).

6 Conclusions and future work

In this work, we conducted a controlled eye-tracking experiment, in which we aimed to predict users' Big-Five personality traits using data captured as part of their interaction with typical RS interfaces. We note four key findings: (1) Gini-Index (GI) score-based feature selector performs more effectively than other selection methods; (2) AdaBoost (AB) combined with GI predicts the traits most accurately, while Decision Tree, Gradient Boosting Decision Tree, and Random Forest offer solid alternatives; (3) The accuracy of personality predictions varies across recommendation interfaces and application domains; and (4) Interface- and domain-specific data allow to improve the accuracy of personality trait predictions. We believe that our results pave the way to the development of future unobtrusive approaches for personality acquisition and modeling, which have the potential to increase the practical applicability of personality-based recommender systems.

In the future, we plan to conduct more studies to verify our findings for other populations, e.g., people from diverse cultural backgrounds, age groups, education levels, and professions, as well as for clinical populations and subjects with personality disorders. Moreover, we are eager to combine the eye-tracking sensors deployed in this work with other technologies, such as face trackers and EEG, to validate the possibility of further increasing the prediction accuracy. Importantly, we would like to have our work deployed in a real-life recommender system, to assess its contribution for practical recommendation metrics, e.g., uptake of recommendations, usability, and potentially system profitability. The ethical aspects related to such a personality detection method should not be under-estimated and also need to be investigated, so as to identify the ways to mitigate the biases they may cause.

Acknowledgements This work was supported by Hong Kong Research Grants Council (project RGC/HKB U12201620) and partially by Hong Kong Baptist University (IRCMS Project IRCMS/19-20/D05). We also thank all participants for their time in taking part in our experiment and reviewers for their valuable comments on our manuscript.

Interface screenshots

See Figs. 7 and 8.

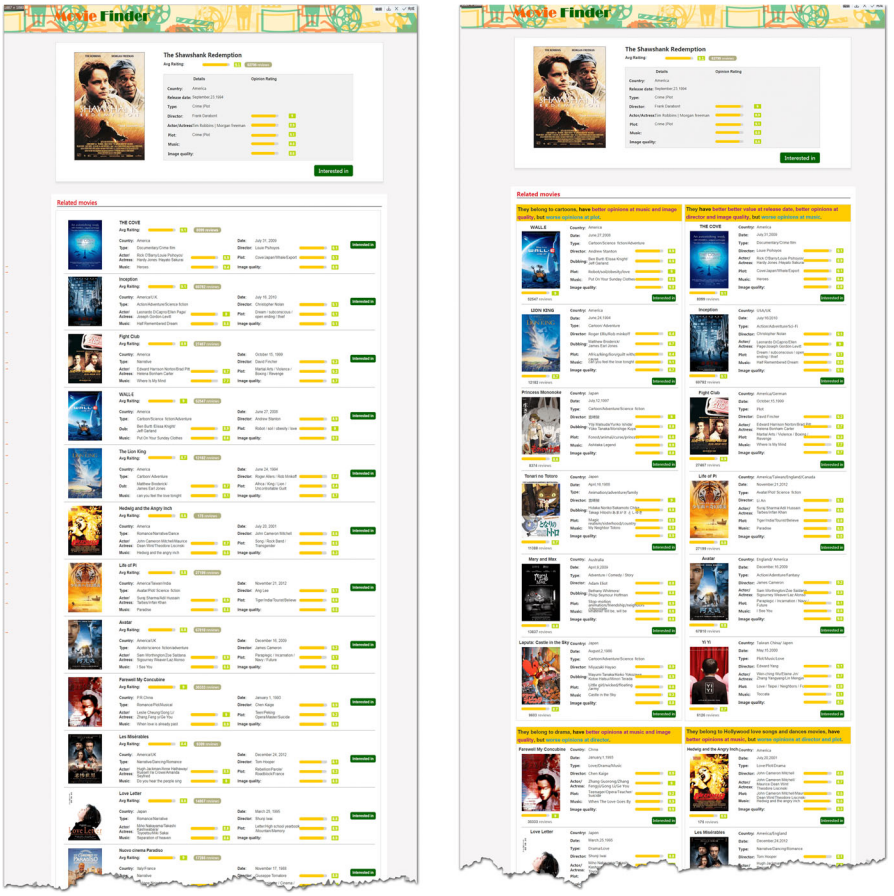


Fig. 7 LIST interface (left) and ORG interface (right) for movies

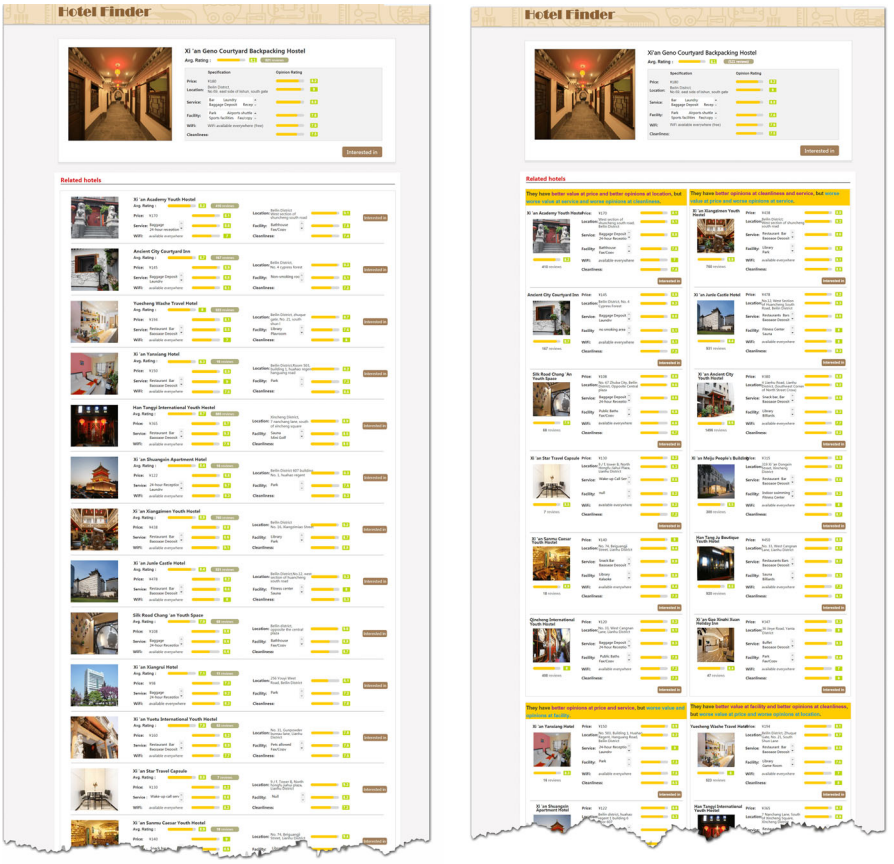


Fig. 8 LIST interface (left) and ORG interface (right) for hotels

References

Ajzen, I.: Attitudes, Personality, and Behavior. McGraw-Hill Education, Bershire (2005)

Alves, T., Natálio, J., Henriques-Calado, J., Gama, S.: Incorporating personality in user interface design: a review. Personal. Individ. Differ. **155**, 109709 (2020)

Anglim, J., Bozic, S., Little, J., Lievens, F.: Response distortion on personality tests in applicants: comparing high-stakes to low-stakes medical settings. Adv. Health Sci. Educ. **23**, 311–321 (2018)

Ashby, N.J.S., Johnson, J.G., Krajbich, I., Wedel, M.: Applications and innovations of eye-movement research in judgment and decision making. J. Behav. Decis. Mak. **29**(2–3), 96–102 (2016)

Ashby, W.L.G.A.N.J.: The effect of consumer ratings and attentional allocation on product valuations. Judgm. Decis. Mak. **10**(2), 172–184 (2015)

Azucar, D., Marengo, D., Settanni, M.: Predicting the big 5 personality traits from digital footprints on social media: a meta-analysis. Personal. Individ. Differ. **124**, 150–159 (2018)

Berkovsky, S., Taib, R., Koprinska, I., Wang, E., Zeng, Y., Li, J., Kleitman, S.: Detecting personality traits using eye-tracking data. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–12 (2019)

Bott, N.T., Lange, A., Rentz, D., Buffalo, E., Clopton, P., Zola, S.: Web camera based eye tracking to assess visual memory on a visual paired comparison task. Front. Neurosci. **11**, 370 (2017)

Cantador, I., Fernández-tobías, I., Bellogín, A.: Relating personality types with user preferences in multiple entertainment domains. In: EMPIRE 1st Workshop on Emotions and Personality in Personalized Services (2013)

Cavanagh, J.F., Wiecki, T.V., Kochar, A., Frank, M.: Eye tracking and pupillometry are indicators of dissociable latent decision processes. J. Exp. Psychol. **143**(4), 1476–1488 (2014)

Chen, F., Ruiz, N., Choi, E., Epps, J., Khawaja, M.A., Taib, R., Yin, B., Wang, Y.: Multimodal behavior and interaction as indicators of cognitive load. ACM Trans. Interact. Intell. Syst. **2**(4), 1–36 (2013)

Chen, F., Ruiz, N., Choi, E., Epps, J., Khawaja, M.A., Taib, R., Yin, B., Wang, Y.: Multimodal behavior and interaction as indicators of cognitive load. ACM Trans. Interact. Intell. Syst. **2**(4), 22:1-22:36 (2013)

Chen, L.: Towards three-stage recommender support for online consumers: implications from a user study. In: International Conference on Web Information Systems Engineering, pp. 365–375 (2010)

Chen, L., Pu, P.: Experiments on the preference-based organization interface in recommender systems. ACM Trans. Comput. Hum. Interact. **17**(1), 1–33 (2010)

Chen, L., Pu, P.: Eye-tracking study of user behavior in recommender interfaces. In: International Conference on User Modeling, Adaptation, and Personalization, pp. 375–380 (2010b)

Chen, L., Pu, P.: Users' eye gaze pattern in organization-based recommender interfaces. In: Proceedings of the 16th International Conference on Intelligent User Interfaces, pp. 311–314 (2011)

Chen, L., Pu, P.: Experiments on user experiences with recommender interfaces. Behav. Inf. Technol. **33**(4), 372–394 (2014)

Chen, L., Wang, F.: Explaining recommendations based on feature sentiments in product reviews. In: Proceedings of the 22nd International Conference on Intelligent User Interfaces, pp. 17–28 (2017)

Chen, L., Wu, W., He, L.: How personality influences users' needs for recommendation diversity? In: CHI'13 Extended Abstracts on Human Factors in Computing Systems, pp. 829–834 (2013c)

Carciofo, R., Yang, J., Song, N., Du, F., Zhang, K: Psychometric evaluation of Chinese-language 44-item and 10-item big five personality inventories, including correlations with chronotype, mindfulness and mind wandering. PLoS ONE **11**(2): e0149963 (2016)

Chen, L., Yan, D., Wang, F.: User evaluations on sentiment-based recommendation explanations. ACM Trans. Interact. Intell. Syst. **9**(4), 1–38 (2019)

Chittaranjan, G., Blom, J., Gatica-Perez, D.: Mining large-scale smartphone data for personality studies. Pers. Ubiquit. Comput. **17**(3), 433–450 (2011)

Costa, P.T., McCrae, R.R.: Neo Personality Inventory-Revised (NEO PI-R). Psychological Assessment Resources Odessa, FL (1992)

Dumais, S.T., Buscher, G., Cutrell, E.: Individual differences in gaze patterns for web search. In: Proceedings of the Third Symposium on Information Interaction in Context, pp. 185–194 (2010)

Elahi, M., Braunhofer, M., Ricci, F., Tkalcic, M.: Personality-based active learning for collaborative filtering recommender systems. In: Congress of the Italian Association for Artificial Intelligence, pp 360–371 (2013)

Fahey, G.: Faking good and personality assessments of job applicants: a review of the literature. DBS Bus. Rev. **2**, 45–68 (2018)

Fernández-Tobías, I., Braunhofer, M., Elahi, M., Ricci, F., Cantador, I.: Alleviating the new user problem in collaborative filtering by exploiting personality information. User Model. User Adapt. Interact. **26**(2–3), 221–255 (2016)

Ferwerda, B., Schedl, M., Tkalcic, M.: Predicting personality traits with instagram pictures. In: Proceedings of the 3rd Workshop on Emotions and Personality in Personalized Systems, pp. 7–10 (2015)

Ferwerda, B., Graus, M.P., Vall, A., Tkalcic, M., Schedl, M.: The influence of users' personality traits on satisfaction and attractiveness of diversified recommendation lists. In: Proceedings of the 4th Workshop on Emotions and Personality in Personalized Systems co-located with ACM Conference on Recommender Systems, pp. 43–47 (2016)

Franco-Watkins, A.M., Johnson, J.G.: Decision moving window: using interactive eye tracking to examine decision processes. Behav. Res. Methods **43**(853), 329–358 (2011)

Gao, R., Hao, B., Bai, S., Li, L., Li, A., Zhu, T.: Improving user profile with personality traits predicted from social media content. In: Proceedings of the 7th ACM Conference on Recommender Systems, pp. 355–358 (2013)

Glaholt, M.G., Reingold, E.M.: Eye movement monitoring as a process tracing methodology in decision making research. J. Neurosci. Psychol. Econ. **4**(2), 125–146 (2011)

Glöckner, A., Herbold, A.K.: An eye-tracking study on information processing in risky decisions: evidence for compensatory strategies based on automatic processes. J. Behav. Decis. Mak. **24**(1), 71–98 (2011)

Golbandi, N., Koren, Y., Lempel, R.: On bootstrapping recommender systems. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 1805–1808 (2010)
Goldberg, L.R.: An alternative "description of personality": the big-five factor structure. J. Pers. Soc. Psychol. **59**(6), 1216–1229 (1990)
Goldberg, L.R., Johnson, J.A., Eber, H.W., Hogan, R., Ashton, M.C., Cloninger, C.R., Gough, H.G.: The international personality item pool and the future of public-domain personality measures. J. Res. Pers. **40**(1), 84–96 (2006)
Gosling, S.D., Rentfrow, P.J., Swann, W.B.: A very brief measure of the big-five personality domains. J. Res. Pers. **37**(6), 504–528 (2003)
Hoppe, S., Loetscher, T., Morey, S.A., Bulling, A.: Eye movements during everyday behavior predict personality traits. Front. Hum. Neurosci. **12**(1), 105 (2018)
Howell, D.C.: Statistical methods for psychology. Cengage Learning (2012)
Hu, R., Pu, P.: A study on user perception of personality-based recommender systems. In: International Conference on User Modeling, Adaptation, and Personalization, pp. 291–302 (2010a)
Hu, R., Pu, P.: Using personality information in collaborative filtering for new users. In: The 2nd Workshop on Recommender Systems and the Social Web co-located with ACM Conference on Recommender Systems, pp. 17–24 (2010b)
Hu, R., Pu, P.: Enhancing recommendation diversity with organization interfaces. In: Proceedings of the 16th International Conference on Intelligent User Interfaces, pp. 347–350 (2011)
Hu, R., Pu, P.: Exploring relations between personality and user rating behaviors. In: The 1st Workshop on Emotions and Personality in Personalized Services co-located with ACM Conference on User Modeling, Adaptation, and Personalization, pp. 1–12 (2013)
Iacobucci, D., Posavac, S.S., Kardes, F.R., Schneider, M.J., Popovich, D.L.: The median split: robust, refined, and revived. J. Consum. Psychol. **25**(4), 690–704 (2015)
John, O.P., Srivastava, S., et al.: The big five trait taxonomy: history, measurement, and theoretical perspectives. Handb. Person. Theory Res. **2**(1999), 102–138 (1999)
Karumur, R.P., Nguyen, T.T., Konstan, J.A.: Personality, user preferences and behavior in recommender systems. Inf. Syst. Front. **20**(6), 1241–1265 (2018)
Kim, B., Park, J., Suh, J.: Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. Decis. Support Syst. **134**, 113302 (2020)
Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. Proc. Natl. Acad. Sci. 110(15), 5802–5805 (2013)
Kret, S.S.E.M.E.: Preprocessing pupil size data: guidelines and code. Behav. Res. Methods **51**, 1336–1342 (2019)
Lancry-Dayan, O.C., Nahari, T., Ben-Shakhar, G., Pertzov, Y.: Do you know him? Gaze dynamics toward familiar faces on a concealed information test. J. Appl. Res. Mem. Cogn. **7**(2), 291–302 (2018)
Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., Liu, H.: Feature selection: a data perspective. ACM Comput. Surv. **50**(6), 1–45 (2017)
Li, L., Li, A., Hao, B., Guan, Z., Zhu, T.: Predicting active users' personality based on micro-blogging behaviors. PLoS ONE **9**(1), e84997 (2014)
Lim, K.K., Friedrich, M., Radun, J., Jokinen, K.: Lying through the eyes: detecting lies through eye movements. In: Proceedings of the Workshop on Eye gaze in Intelligent Human Machine Interaction: Gaze in Multimodal Interaction, pp. 51–56 (2013)
Lu, J., Wu, D., Mao, M., Wang, W., Zhang, G.: Recommender system application developments: a survey. Decis. Support Syst. **74**(1), 12–32 (2015)
Majumder, N., Poria, S., Gelbukh, A., Cambria, E.: Deep learning-based document modeling for personality detection from text. IEEE Intell. Syst. **32**(2), 74–79 (2017)
Manolios, S., Hanjalic, A., Liem, C.C.S.: The influence of personal values on music taste. In: Proceedings of the 13th ACM Conference on Recommender Systems, pp. 501–505 (2019)
McClendon, J., Bogdan, R., Jackson, J.J., Oltmanns, T.F.: Mechanisms of black-white disparities in health among older adults: examining discrimination and personality. J. Health Psychol. 26(7), 995–1011 (2019)
McCrae, R.R., Costa Jr, P.T.: Conceptions and correlates of openness to experience. In: Handbook of Personality Psychology, pp. 825–847 (1997)
McCrae, R.R., John, O.P.: An introduction to the five-factor model and its applications. J. Pers. **60**(2), 175–215 (1992)

Millecamp, M., Htun, N.N., Conati, C., Verbert, K.: What's in a user? towards personalising transparency for music recommender interfaces. In: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, Association for Computing Machinery, New York, NY, USA, UMAP '20, pp. 173-182 (2020)

Millecamp, M., Conati, C., Verbert, K.: Classifeye: Classification of personal characteristics based on eye tracking data in a recommender system interface. In: Joint Proceedings of the ACM IUI 2021 Workshops (2021)

Mitsuda, T., Glaholt, M.G.: Gaze bias during visual preference judgements: effects of stimulus category and decision instructions. Vis. Cogn. **22**(1), 11–29 (2014)

Morey, L.C., Gunderson, J., Quigley, B.D., Lyons, M.: Dimensions and categories: the "big five" factors and the DSM personality disorders. Assessment **7**(3), 203–216 (2000)

Mounica, M.S., Manvita, M., Jyotsna, C., Amudha, J.: Low cost eye gaze tracker using web camera. In: 3rd International Conference on Computing Methodologies and Communication, pp. 79–85 (2019)

Nguyen, T.T., Harper, F.M., Terveen, L., Konstan, J.A.: User personality and user satisfaction with recommender systems. Inf. Syst. Front. **20**(6), 1173–1189 (2018)

Nicholson, N., Soane, E., Fenton-O'Creevy, M., Willman, P.: Personality and domain-specific risk taking. J. Risk Res. **8**(2), 157–176 (2005)

Pachur, T., Spaar, M.: Domain-specific preferences for intuition and deliberation in decision making. J. Appl. Res. Mem. Cogn. **4**(3), 303–311 (2015)

Poole, A., Ball, L.J.: Eye tracking in human–computer interaction and usability research: Current status and future. In: Encyclopedia of Human–Computer Interaction, pp. 211–219 (2005)

Poropat, A.E.: A meta-analysis of the five-factor model of personality and academic performance. Psychol. Bull. **135**(2), 322 (2009)

Pu, P., Chen, L.: Trust building with explanation interfaces. In: Proceedings of the 11th International Conference on Intelligent User Interfaces, pp. 93–100 (2006)

Pu, P., Chen, L.: Trust-inspiring explanation interfaces for recommender systems. Knowl. Based Syst. **20**(6), 542–556 (2007)

Purvis, A., Howell, R.T., Iyer, R.: Exploring the role of personality in the relationship between maximization and well-being. Person. Individ. Differ. **50**(3), 370–375 (2011)

Quercia, D., Kosinski, M., Stillwell, D., Crowcroft, J.: Our twitter profiles, our selves: predicting personality with twitter. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, pp. 180–185 (2011)

Raptis, G.E., Fidas, C.A., Avouris, N.M.: On implicit elicitation of cognitive strategies using gaze transition entropies in pattern recognition tasks. In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, pp. 1993–2000 (2017)

Rauthmann, J.F., Seubert, C.T., Sachse, P., Furtner, M.R.: Eyes as windows to the soul: gazing behavior is related to personality. J. Res. Pers. **46**(2), 147–156 (2012)

Rayner, K.: Eye movements in reading and information processing: 20 years of research. Psychol. Bull. **124**(3), 372–422 (1998)

Rentfrow, P., Gosling, S.: The do re mi's of everyday life: the structure and personality correlates of music preferences. J. Personal. Soc. Psychol. **84**(6), 1236–1256 (2003)

Riaz, M.N., Riaz, M.A., Batool, N.: Personality types as predictors of decision making styles. J. Behav. Sci. **22**(2), 99–114 (2012)

Ricci, F., Rokach, L., Shapira, B.: Recommender Systems Handbook, 2nd edn. Springer Publishing Company, (2015)

Rojas, J.C., Marín-Morales, J., Ausín Azofra, J.M., Contero, M.: Recognizing decision-making using eye movement: a case study with children. Front. Psychol. **11**, 2542 (2020)

Sadi, R., Asl, H.G., Rostami, M.R., Gholipour, A., Gholipour, F.: Behavioral finance: the explanation of investors' personality and perceptual biases effects on financial decisions. Int. J. Econ. Financ. **3**(5), 234–241 (2011)

Salvucci, D.D., Goldberg, J.H.: Identifying fixations and saccades in eye-tracking protocols. In: Proceedings of the 2000 Symposium on Eye Tracking Research & Applications, pp. 71–78 (2000)

Shahjehan, A., Zeb, F., Saifullah, K., et al.: The effect of personality on impulsive and compulsive buying behaviors. Afr. J. Bus. Manag. **6**(6), 2187–2194 (2012)

Sharan, R.V., Berkovsky, S., Taib, R., Koprinska, I., Li, J.: Detecting personality traits using interhemispheric asynchrony of the brainwaves. In: 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, pp. 62–65 (2020)

Shen, J., Brdiczka, O., Liu, J.: Understanding email writers: personality prediction from email messages. In: User Modeling, Adaptation, and Personalization, pp. 318–330 (2013)
Stewart, N., Hermens, F., Matthews, W.J.: Eye movements in risky choice. J. Behav. Decis. Mak. **29**(2–3), 116–136 (2016)
Stoeber, J., Otto, K., Dalbert, C.: Perfectionism and the big five: conscientiousness predicts longitudinal increases in self-oriented perfectionism. Personal. Individ. Differ. **47**(4), 363–368 (2009)
Tai, R.H., Loehr, J.F., Brigham, F.J.: An exploration of the use of eye-gaze tracking to study problem-solving on standardized science assessments. Int. J. Res. Method Educ. **29**(2), 185–208 (2006)
Taib, R., Berkovsky, S., Koprinska, I., Wang, E., Zeng, Y., Li, J.: Personality sensing: detection of personality traits using physiological responses to image and video stimuli. ACM Trans. Interact. Intell. Syst. **10**(3), 181–1832 (2020)
Tintarev, N., Masthoff, J.: Evaluating the effectiveness of explanations for recommender systems. User Model. User Adapt. Interact. **22**(4–5), 399–439 (2012)
Tintarev, N., Dennis, M., Masthoff, J.: Adapting recommendation diversity to openness to experience: a study of human behaviour. In: International Conference on User Modeling, Adaptation, and Personalization, pp. 190–202 (2013)
Tiwari, V., Ashpilaya, A., Vedita, P., Daripa, U., Paltani, P.P.: Exploring demographics and personality traits in recommendation system to address cold start problem. pp. 361–369 (2020)
Tkalcic, M., Chen, L.: Personality and recommender systems. In: Recommender Systems Handbook, pp. 715–739 (2015)
Tkalcic, M., Kunaver, M., Tasic, J., Košir, A.: Personality based user similarity measure for a collaborative recommender system. In: Proceedings of the 5th Workshop on Emotion in Human–Computer Interaction-Real world challenges, pp. 30–37 (2009)
Tkalcic, M., Quercia, D., Graf, S.: Preface to the special issue on personality in personalized systems. User Model. User Adapt. Interact. **26**(2–3), 103–107 (2016)
Toker, D., Conati, C., Carenini, G.: Gaze analysis of user characteristics in magazine style narrative visualizations. User Model. User Adapt. Interact. **29**, 1011–977 (2019)
Valtakari, N.V., Hooge, I.T.C., Viktorsson, C., Nyström, P., Falck-Ytter, T., Hessels, R.S.: Eye tracking in human interaction: Possibilities and limitations. In: Companion Publication of the 2020 International Conference on Multimodal Interaction, p. 508 (2020)
Van Lankveld, G., Spronck, P., Van den Herik, J., Arntz, A.: Games as personality profiling tools. In: 2011 IEEE Conference on Computational Intelligence and Games, pp. 197–202 (2011)
Van Nuenen, T., Ferrer, X., Such, J.M., Cote, M.: Transparency for whom? Assessing discriminatory artificial intelligence. Computer **53**(11), 36–44 (2020)
Wache, J., Subramanian, R., Abadi, M.K., Vieriu, R.L., Sebe, N., Winkler, S.: Implicit user-centric personality recognition based on physiological responses to emotional videos. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 239–246 (2015)
Wang, K., Ji, Q.: Real time eye gaze tracking with 3D deformable eye-face model. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1003–1011 (2017)
Wilbers, A.K., Vennekoetter, A., Kacauster, M., Hamborg, K.C., Kaspar, K.: (2015) Personality traits and eye movements: an eye-tracking and pupillometry study. In: Proceddings of the European Conference on Eye Movement, p. 268
Wu, W., Chen, L.: Implicit acquisition of user personality for augmenting movie recommendations. In: International Conference on User Modeling, Adaptation, and Personalization, Springer, pp. 302–314 (2015)
Wu, W., Chen, L., Zhao, Y.: Personalizing recommendation diversity based on user personality. User Model. User Adapt. Interact. **28**(3), 237–276 (2018)
Xu, J., Wang, Y., Chen, F., Choi, E.: Pupillary response based cognitive workload measurement under luminance changes. In: IFIP Conference on Human–Computer Interaction, pp. 178–185 (2011)
Zhang, X., Liu, X., Yuan, S.M., Lin, S.F., Mehmood, I.: Eye tracking based control system for natural human-computer interaction. Computational Intelligence and Neuroscience (2017)
Ziegler, M., MacCann, C., Roberts, R.: New perspectives on faking in personality assessment (2011)
Zillig, L.M.P., Hemenover, S.H., Dienstbier, R.A.: What do we assess when we assess a Big 5 trait? A content analysis of the affective, behavioral, and cognitive processes represented in Big 5 personality inventories. Pers. Soc. Psychol. Bull. **28**(6), 847–858 (2002)

Li Chen is an associate professor in the Department of Computer Science at Hong Kong Baptist University, China. Her research interests are mainly in the areas of human-centered AI, recommender systems, and intelligent user interfaces.

Wanling Cai is a PhD candidate in the Department of Computer Science at Hong Kong Baptist University, under the supervision of Dr. Li Chen. Her research mainly focuses on human-centered design for conversational agents and recommender systems.

Dongning Yan is a lecturer in the Industrial Design Institute at Shandong University, China. She received her PhD in Design from Politecnico di Milano in 2015. Her research focuses are on user-centered design and user experience research for human–product interaction.

Shlomo Berkovsky is a professor at the Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University. His core research areas are user modeling, personalization, and recommender systems, which lie on the intersection of data science and human–computer interaction.