Trusting Virtual Agents: The Effect of Personality

MICHELLE X. ZHOU, Juji, Inc.

GLORIA MARK, University of California, Irvine

JINGYI LI, University of California, Berkeley

HUAHAI YANG, Juji, Inc.

We present artificial intelligent (AI) agents that act as interviewers to engage with a user in a text-based conversation and automatically infer the user's personality traits. We investigate how the personality of an AI interviewer and the inferred personality of a user influences the user's trust in the AI interviewer from two perspectives: the user's willingness to confide in and listen to an AI interviewer. We have developed two AI interviewers with distinct personalities and deployed them in a series of real-world events. We present findings from four such deployments involving 1,280 users, including 606 actual job applicants. Notably, users are more willing to confide in and listen to an AI interviewer with a serious, assertive personality in a high-stakes job interview. Moreover, users' personality traits, inferred from their chat text, along with interview context, influence their perception of and their willingness to confide in and listen to an AI interviewer. Finally, we discuss the design implications of our work on building hyper-personalized, intelligent agents.

CCS Concepts: • **Human-centered computing** → **Interactive systems and tools**; • **Computing methodologies** → **Intelligent agents**;

Additional Key Words and Phrases: AI interviewer, conversational agents, chatbots, personality inference, human-machine trust, computer personality, individual differences, Big 5 personality

ACM Reference format:

Michelle X. Zhou, Gloria Mark, Jingyi Li, and Huahai Yang. 2019. Trusting Virtual Agents: The Effect of Personality. *ACM Trans. Interact. Intell. Syst.* 9, 2–3, Article 10 (March 2019), 36 pages.

<https://doi.org/10.1145/3232077>

1 INTRODUCTION

In a wide variety of domains, including human capital management and patient care, interviewing is often used to collect information from target respondents and to assess certain characteristics of the respondents. For example, a hiring manager may interview a pool of job applicants over the phone to discover their career interests and assess their characteristics, such as personality and motivations, before determining their fit for the position.

While a human interviewer can gather in-depth information, it presents several limitations. First, it is difficult to scale since a human interviewer can interview only one person at a time

The reviewing of this article was managed by special issue associate editors Fang Chen, Carlos Duarte, and Wai-Tat Fu.

This work is funded in part by the Air Force Office of Scientific Research under FA9550-15-C-0032.

Authors' address: M. X. Zhou and H. Yang, Juji, Inc., 3165 Olin Ave., San Jose, 95117; emails: mzhou@acm.org, hyang@juji-inc.com; G. Mark, Dept. of Informatics, University of California, Irvine, CA; email: gmark@uc.edu; J. Li, Dept. of Computer Science, Stanford University, Stanford, CA; email: jingyili@stanford.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2019 Association for Computing Machinery.

2160-6455/2019/03-ART10 \$15.00

<https://doi.org/10.1145/3232077>

Fig. 1. A screenshot of a chat between an AI interviewer (Kaya) and a user (James).

and conduct only a limited number of interviews per day. Second, research shows that a human interviewer's personal factors, such as personality, can affect interview results (e.g., see Refs [45, 106]). Third, respondents may not feel comfortable disclosing sensitive information to human interviewers due to social desirability biases [59].

To overcome the limitations of a human interviewer while providing a personal touch, we have built artificial intelligence (AI) interviewers that conduct one-on-one interviews via text-based conversations. Figure 1 shows a screenshot where an AI interviewer (Kaya) conducts a text-based interview with a user (James). During such an interview, Kaya also *automatically* infers James's characteristics, such as personality traits, based on his chat text given during the interview.

Our decision to support a text-only conversation is based on three considerations. First, with the increasing use of cell phones and social media, more and more people use text-based communication, also known as texting. A Gallup poll¹ shows that texting is the dominant communication method for Americans under 50. We thus wish to support human-computer conversations in a modality that most people are familiar with. Second, an interview is often a private activity, especially in a high-stakes context such as job interviewing. Text-only interaction not only keeps a user focused on the interview without distractions [48, 86], but also lets the user do so discreetly without disturbing others nearby as a voice-based interaction might. Third, although significant technical advances have been made to support human-computer interactions in richer modalities such as voice and gestures (e.g., see Refs [22, 75]), the technologies including those used in commercial systems like Apple Siri, Amazon Alexa, and Google Now are far from being perfect [25, 60, 93, 102]. Given our goal of deploying AI interviewers for real-world use, we decided to support text-only conversations to minimize potential system errors due to technology imperfections (e.g., speech recognition errors due to user accent or a noisy surrounding).

¹<http://news.gallup.com/poll/179288/new-era-communication-americans.aspx>.

Compared to a human interviewer, an AI interviewer has two distinct advantages. First, it does not experience fatigue and can scale to interview hundreds of thousands of users simultaneously. Second, it can render a more objective, unbiased assessment, since it does not bring personal emotions or human prejudice into an interview as a human interviewer might (e.g., affect heuristic [107]).

Since studies show that interviewers' personalities directly affect interview outcomes [45, 106], one of our tasks is to equip an AI interviewer with the "personality" of an effective interviewer. This is non-trivial for three reasons. First, existing research suggests that computers should take on *one* strong, consistent personality [78]. However, it is unclear which personality an AI interviewer should assume in order to achieve multiple interview goals. For example, a warm and cheerful personality may aid in building rapport with human respondents who can then relax and act authentically [9, 59]. On the other hand, a more serious and assertive personality may elicit more respect and better cooperation from the respondents [45, 77]. Second, existing systems stress the use of both verbal and non-verbal cues to express an agent's personality [54, 87, 77]. However, it is unclear how an AI interviewer can effectively exhibit its personality through text-only expressions. Third, unlike most AI agents, which are used in a laboratory setting, ours will be used in the real world, including high-stakes situations, such as job interviews. It is unknown how an agent's personality would affect users in such a context.

To address the challenges mentioned above, as the first step, we decided to give each AI interviewer a strong, consistent personality as suggested by existing work [78]. We built two AI interviewers - called Responsible and Empathetic Personas (REPs), with two distinct personalities, each aimed at achieving a key interview goal. One is a female persona (Kaya) with a warm and cheerful personality, while the other is a male persona (Albert) with a reserved and assertive personality. Based on existing work and through various deployments of Kaya and Albert, we hope to answer three research questions:

- **R1:** How does the personality of an agent impact user's willingness to confide in that REP?
- **R2:** How does the personality of an agent impact users' willingness to listen to that REP?
- **R3:** How do other factors, such as interview context and users' individual differences, including personality traits, impact their behavior with an agent?

We have deployed our AI interviewers for a number of uses. Here, we use the results of four real-world deployments to answer research questions mentioned above. In all four deployments, our AI interviewers asked a series of questions to collect information from target respondents and inferred the respondents' Big 5 personality traits—the most widely used personality model [23] based on their responses. The first deployment used both Kaya and Albert to help an organization interview and pre-screen job applicants. In this deployment, 157 applicants finished their interview with Kaya, and 159 completed theirs with Albert. Another three deployments all used Kaya to interview target respondents for varied purposes: (a) interviewing and pre-screening summer internship applicants (290 completed interviews), (b) interviewing university students to elicit their team experience and preferences and infer their characteristics to facilitate project teaming (201 completed interviews), and (c) interviewing university students to assess student characteristics for educational purposes (473 completed interviews).

As seen later, our work offers two unique contributions. First, we present novel, practical approaches to building effective AI interviewers for real-world uses. In particular, we have built a topic-based conversation engine that can reliably support the completion of hour-long structured interviews. Moreover, our approach allows easy and fast customization of an AI interviewer to suit diverse interview scenarios, which helps the real-world adoption of such agents. Our AI interviewers can also dynamically analyze users' responses to automatically assess their characteristics

on the fly and incorporate the inferred user characteristics into the interview to make an interview more engaging.

Second, the results of our analyses of the four deployments reveal several new insights in understanding users' trust in intelligent agents in real-world settings. Specifically, the personality of an agent affects users' willingness to confide in and listen to the REP. Furthermore, users' personality traits inferred by an agent influence their perception of and behavior with the agent. As a result, these findings inform design implications for building hyper-personalized agents.

2 RELATED WORK

Our work is built upon numerous existing efforts on studying and building effective virtual agents across multiple disciplines, from Psychology to Communications to Computer Science. These related efforts broadly fall in the following areas.

2.1 Computer Anthropomorphism and User Behavior

Our work builds AI interviewers that can influence and guide a user's behavior (e.g., truth telling). It is directly related to a rich body of research on examining how humans respond to computers with varied human-like characteristics. One line of studies focuses on how humans apply similar social rules used in the real world to computers as social actors. For example, Nass and his colleagues show that computer agents elicit users' social attributions, which are different based on the characteristics of the agents (e.g., the type of computer voice used) [69, 78]. In another example, Fogg and Nass use a laboratory experiment to show that humans enjoy the flattery from computers as much as from other humans [28].

In addition to understanding human social behavior around computers, another extensive line of work examines humans' responses to computers with more expressive, human-like qualities, such as faces and facial expressions. In general, these studies have found that anthropomorphic properties of computers influence users' perceptions (e.g., [13, 70, 86]), attitude (e.g., [13]), and behavior around such systems (e.g., [48, 86, 105]). For example, Sproull et al. [86] report that people respond to a text-based interface differently than to a talking face. On the one hand, users are aroused more and present themselves more positively when interacting with a talking face. On the other hand, users are less relaxed or assured when interacting with a talking face. Another study shows that users find the interface with the anthropomorphic qualities—faces and facial expressions—more likeable and engaging, although such an interface takes the users' effort to interpret the meaning of the human-like expressions and may even be a distraction [48, 86]. More recent studies show that human-like features with higher realism elicit more positive social interactions while having no significant impact on user task performance [105]. Furthermore, a study reveals that anthropomorphism may even elicit user objections due to users' own biases (e.g., sexism) [70].

Our work draws upon findings from these existing efforts. For example, we decide to support a text-only conversation to minimize potential user distraction or unnecessary biases rising from interpreting human-like expressions of an agent (e.g., feeling judged). Additionally, we focus on instilling specific intelligence and behavior into our agents so that they can function robustly to accomplish real-world tasks effectively (e.g., job interviews).

2.2 Computer Personality and User Personality

Human users often attribute personality to a machine with anthropomorphic characteristics. To understand how user-perceived machine personality influences user behavior, there are numerous efforts investigating the relationships among machine personality, user personality, and tasks (e.g., see Refs [13, 24, 54, 68, 87]). For example, multiple studies find that users favor a computer with a strong, consistent personality [68, 87]. Moreover, users favor machines with a personality that

matches their own personality [68]. However, in human-robot interaction, studies show conflicting results. For example, Lee et al. [54] indicate that users enjoy interacting with a robot with a complementary personality, while Tapus et al. [87] show that users prefer to interact with a robot with a matching personality. Moreover, studies find that users expect the personality of an agent to match the task context, e.g., a serious agent for a serious task [77], which directly influences users' willingness to comply with the agent's instructions.

Existing studies provide insights into the effects of an agent's personality on user behavior in various laboratory situations. However, it is unknown what personality an AI interviewer should take on and how its personality would impact its interviewees, especially in real-world situations, including high-stakes contexts like job interviews. Therefore, our work aims at answering this line of questions.

2.3 Intelligent Virtual Agents for Interviewing

Our effort of building AI interviewers is directly related to a line of work on developing embodied "virtual human interviewers" (e.g., see Refs [22, 30, 46, 59, 71]). Closest to our effort, DeVault et al. [22] built a virtual human interviewer to conduct interviews and assess distress indicators of users by analyzing their verbal and non-verbal behavior. Similarly, Nunamaker et al. have developed an information kiosk that acts as a conversational agent to interview users and detect their affective state and behavior. While our goals are similar in evaluating users' characteristics through an AI interview, our focus is different. Specifically, our work analyzes a user's communication behavior to infer the user's innate psychological characteristics such as personality, while theirs is on detecting changes in one's behavior during the interaction (e.g., arousal or distress).

2.4 Relational Agents

Since our work studies users' trust in virtual agents, it is related to numerous efforts on building trustworthy, relational agents (e.g., see Refs [9, 10, 46]). For example, researchers have built relational agents to interact with patients to encourage behavioral change [9] or coach a user on her/his interview behavior [46].

Our current effort is built on prior work of building relational agents. For example, our AI interviewer uses introductory questions [59] and reciprocal self-disclosure [9] in relational agents to build rapport with a user, and adopts different agents' personas to influence users' behavior [30, 46]. However, our work differs in several aspects. First, since most relational agents are built to perform coaching or counseling tasks, existing efforts focus on instrumenting social skills of such an agent to build connections with users. In contrast, the primary purpose of ours is to conduct interviews and we focus on their ability to effectively elicit information in a conversation. Second, unlike most relational agents that employ both verbal and non-verbal cues to exhibit their personas, ours must present a strong, consistent personality via text-only communication. As discussed later, our work addresses unique challenges that have not been addressed before, which in turn makes unique contributions to building effective AI interviewers for real-world tasks.

2.5 Synthesis of Machine Personality

Our work on instilling a personality into an AI interviewer is related to various efforts on automated synthesis of machine personality. As indicated by a research survey on machine personality synthesis [95], there are three types of effort in general. One type of effort is to exhibit a machine's personality through synthetic voice by manipulating various prosodic features, such as pitch and speaking rate (e.g., see Refs [82, 92]). The second type is to display a virtual agent's personality through various agent properties, including physical appearance, body movement, and language use (e.g., see Refs [5, 65]). The third type is to instrument a physical robot's personality through

its interaction behavior with users, such as speech, gaze, and physical distance to users (e.g., see Refs [3, 87]).

In addition to these efforts, there are theories on building dialog agents with a personality through verbal communication [11, 81]. Since our virtual interviewer engages a user in a text-based communication, our current work is closest to the efforts on using various verbal cues to display machine personality [11, 81]. While we leverage the findings in such work, we also extend these efforts to create machine personality through additional behavior specific to interviewing (e.g., questioning and response style).

2.6 Inference of User Personality

Our work on automatically inferring a user's personality traits from the user's text communication during a virtual interview is inspired by numerous recent efforts on studying the relationships between users' traits and their communication behavior (e.g., see Refs [63, 88, 104, 47, 49, 76]). For example, research indicates that words or word categories in users' blogs and essays are correlated with self-reported or third-party-rated personality traits [63, 88, 104]. More recent studies show that users' vocabulary and social media behavior, such as Facebook and Twitter, are also related to self-reported personality traits [47, 49, 76]. In addition to words, word categories, and social media behavior, researchers have also examined how voices (e.g., see Refs [44, 56]), images (e.g., see Refs [2, 58]), and videos (e.g., see Refs [8, 56]) are related to personality traits.

Closest to our current effort is on computational models to automatically infer one's personality traits from the text generated by the person (e.g., see Refs [1, 35, 62]). Broadly, there are two main types of computational models [95]. One type uses various classification algorithms to determine whether one's personality trait falls into a particular category (e.g., high or low Extroversion) (e.g., see Refs [4, 31]). For example, Linguistic Inquiry and Word Count (LIWC)² word categories are used to classify whether a blogger is "high," "medium," or "low" on each of the Big 5 Personality traits based on the words used in blogs [31]. The other type of effort is to use various machine learning methods, mainly regression algorithms, to predict one's personality trait scores based on various user behavior (e.g., see Refs [51, 76]). For example, a recent study uses regression algorithms to predict one's Dark Triad³ personality scores using a combination of Twitter content, profile image, and usage behavior (e.g., number of followers) [76].

Similar to the second type of computational models, ours automatically infers a user's personality trait scores from text. Existing models infer user traits by correlating users' self-reported psychometric test scores with their text footprints (e.g., see Refs [1, 35, 51, 62]). In contrast, we gather "ground-truth" data (Section 4.1, step 1) to train our model because gathering self-reported scores is expensive and may not be authentic. As a result, our approach is more objective and cost-efficient, since it no longer requires a large sample of self-reported psychometric scores.

2.7 Practical Approaches to Conversational Agents

There are numerous computational approaches to conversational agents, including rule-based and data-driven (deep learning) approaches [91]. Given our goal of building agents for real-world uses, our work is most related to numerous efforts on creating commercial grade conversational agents, such as Amazon Alexa, Google Allo, Microsoft Cortana, and Apple Siri. Most of these

²The LIWC dictionary contains a set of words and word stems. Each word or word stems falls in one or more word categories, such as Pronouns and Numbers, which can be used for text analysis: https://s3-us-west-2.amazonaws.com/downloads.liwc.net/LIWC2015_OperatorManual.pdf.

³Three malevolent personal qualities: *Narcissism*, *Machiavellianism*, and *Psychopathy*. See more here: https://en.wikipedia.org/wiki/Dark_triad.

Fig. 2. System overview.

conversational agents serve as a user's personal assistant to perform specific tasks, such as making a call or controlling a home device, based on the user's verbal requests. Most task-oriented agents are powered by rules [91]. Recently, deep learning techniques are used to train retrieval models for social agents, where the most probable response is retrieved from a repository of pre-defined responses given a user's request [103]. Alternatively, researchers build generative models (e.g., see Refs [96, 84, 85]). While these models can produce machine responses that never existed before, the quality of the generated responses does not work well in practice yet.

Compared to the existing efforts of creating AI agents, our goal is to create AI interviewers who can work well in practice, effectively carrying out a conversation to elicit information from a user and assess the user's characteristics. Neither retrieval-based nor generative models are practical for us, since they require huge amounts of training data [103] and it is difficult to obtain interview data as training data, not mentioning that such training data must cover varied interview contexts and diverse user responses given in an interview [22]. We thus have developed a rule-based approach with algorithms to manage an interview conversation (Section 6).

3 SYSTEM OVERVIEW

Figure 2 shows an overview of our system that supports automated interviews. An AI interviewer engages a user (interviewee) in a text-based conversation (c), powered by two core engines: (a) a personality engine and (b) a conversation engine. The *personality engine* analyzes a user's interaction behavior and automatically infers the user's personality traits and beyond, such as strengths and weaknesses. The *conversation engine* enables an AI interviewer to engage a user in a mixed-initiative conversation. Similar to a human interviewer, an AI interviewer takes initiatives to guide the interview flow and solicit information from a user, while it allows the user to pose questions whenever appropriate. For example, a user may ask a clarification question or inquire about the position that s/he is applying for. In such a case, the *conversation engine* first interprets the user's request and then decides how to respond to the request.

To help business users (e.g., a hiring manager or a market researcher) use interview results to make decisions, we provide an interactive visual dashboard that summarizes interview results (d). Figure 3(a) is a screenshot of such a visual dashboard that shows the number of interviewees who complete an interview and the derived Big 5 personality trait scores for the interviewees. Figure 3(b) displays a more detailed summary for a particular trait, including individuals' scores for that trait. Our system also maintains several databases (DB) and knowledge bases (KB), which store various types of information, such as users' answers to interview questions and inferred user traits.

Fig. 3. A visual dashboard that displays interview results.

4 EVIDENCE-BASED PERSONALITY ENGINE

To automatically infer a user's traits from text, most existing approaches use the uncovered correlations between word categories (i.e., LIWC categories) and self-reported personality trait scores (e.g., see Refs [1, 35, 62]). While these approaches have demonstrated their effectiveness in various applications (e.g., see Refs [55, 61, 62]), they have three main limitations. First, while using word categories is simple, the inference is coarse. This is because different words belonging to the same word category have the same weight, but their relationships to a trait may be quite different [104]. Moreover, there are a limited number of word categories (68 word categories in LIWC), and many word categories may not show significant correlations with a trait. Second, the correlations are derived based on one's word use and self-reported personality test scores, which may not be completely truthful [108]. Third, many of these approaches consider only single words but ignore other important linguistic cues, such as phrases, and punctuations, which may also reflect one's traits.

4.1 Approach

To overcome these limitations, we have developed an evidence-based approach to inferring a user's traits from the user's text. Below we highlight its three main steps.

Step 1: Gathering Trait-Related Evidence

To build a personality inference engine, we first gather training data—user behavioral data that captures people's real-world activities that reflect their associated characteristics. Using the Twitter Application Programming Interface (API), we have collected tweets from 15 million Twitter users, each with up to 3,200 tweets. We then randomly sampled 1 million Twitter users and used their tweets as our training data. There are three reasons why we chose to use Twitter data as our training data. First, Twitter data is publicly available and abundant. Second, not only does Twitter data cover diverse topics, from politics to entertainment, but it is also generated by a diverse population, from retirees to presidents. Such diversity helps avoid potential data biases. Third, existing work, including our own, shows the effectiveness of Twitter data for measuring personality traits (e.g., see Refs [1, 35, 51, 62]).

Given the dataset, tweets were examined and identified as potential trait-related evidence based on existing findings in the literature on personality assessment. For example, DeYoung et al. [20] indicate that a person high on *Artistic Interests* (a facet of *Openness*, one of the Big 5 personality factors) enjoys music and poetry. Based on this finding, we can mark two sample tweets below as evidence reflecting *Artistic Interests*, with one containing reference to poetry and the other to music:

Brian Bilston @brian_bilston · 26 Dec 2017

Here is a nightmarish Christmas poem, entitled 'Upon Awakening to the Sound of Distant Rumbles'.

Lady Gaga @ladygaga · 28 Nov 2017

I'm humbled & grateful that my album "Joanne" was nominated & also my song "Million Reasons." Thank u so much Monsters & Grammy Voters for believing in this music and me. I love u so much. This is a day to truly celebrate lots of hard work, perseverance & gratitude through music.

Based on previous literature, certain evidence can be computed automatically. For example, Costa and McCrae have identified a person high on *Intellect* (another facet of *Openness*) often has a rich vocabulary [17]. Based on such findings, the *richness* of one's vocabulary can be computed automatically based on word use patterns, such as lexical density and diversity, in one's online communications (e.g., see Refs [42, 50]). Based on this finding, a vocabulary richness score can be computed for each tweet and marketed as evidence related to trait *Intellectual Curiosity*.

Since it is impractical to identify each tweet as evidence relevant to one or more personality traits, we identify a Twitter user and associate him/her with one or more traits based on this person's known characteristics ("ground truth"). For example, we associate Lady Gaga (@ladygaga) with *Artistic Interests* given her profession. Likewise, a Stanford math professor (e.g., @joboaler) is related to trait *Intellectual Curiosity*, while a travel enthusiast (@wenphd) is associated with trait *Adventurousness*. Using this method, we identify and collect groups of Twitter users with shared, distinct characteristics (known groups). We then compare each known group's tweets with those of randomly selected users in our training set to extract a set of linguistic cues *unique* to the known group. Since much research shows that linguistic cues reflect personality traits (e.g., see Refs [1, 35, 47, 51]), the extracted linguistic cues, including words, phrases, punctuations, and emoticons, are then considered potential evidence for reflecting the characteristics represented by each known group.

Step 2: Mining Evidence-Trait Relationships

To quantify the relationships between potential evidence (the linguistic cues identified in *Step 1*) and user traits, such as the Big 5 personality traits, we adopt a statistical inference approach, *Item Response Theory*, a widely used approach in modern psychometrics [19]. In particular, we model the desired traits as latent factors and potential evidence as observed items. The following formula can be used to estimate the relationships between evidence j and trait θ (trait score) for an individual i :

$$X_{ij} = \mu_j + \lambda_j \theta_i + \varepsilon_{ij}; \quad j = 1, N. \quad (1)$$

Here, X is the observed data containing N pieces of evidence; μ_j is the occurrence rate of evidence j ; λ_j measures the discriminative power of evidence j to trait θ ; and ε_{ij} is the Gaussian distributed error uncorrelated with θ .

To estimate the parameters in Formula 1, we use the conditional distribution of X for trait θ :

$$f(X_j | \theta_i) = \frac{1}{\sigma_j \sqrt{2\pi}} \frac{1}{X_j(1-X_j)} e^{-\frac{1}{2} \left[\frac{\ln(\frac{X_j}{1-X_j}) - \mu'}{\sigma_j} \right]^2} \quad (2)$$

Here, μ is the expectation of μ after a logit transformation, and σ_j^2 is the variance of residuals. The parameters can then be estimated using various algorithms, such as an Expectation Maximization (EM) algorithm over the training data [21].

Step 3: Evidence-based Inference of Trait Scores

Given a user's text (e.g., Facebook posts, tweets, or chat text), the trained model can be applied to compute trait scores (Formula 1). Currently, we focus on inferring Big 5 personality factors and their 30 associated facets [23]. Any user-generated text (e.g., emails and blogs) can be used to infer a user's personality traits using our model. In all the field deployments described below, a user's chat text collected during an interview is used to infer the user's traits.

4.2 Validating Inferred User Traits

Similar to validating any other psychometric measures (e.g., see Refs [20, 23]), we have evaluated our personality inference on two metrics: *reliability* (how consistent a measure is) and *validity* (how accurate a measure is). To measure the reliability of our model, we used our engine to analyze the tweets of 100,000 randomly sampled users to compute the Cronbach's α , a metric estimating the reliability of a psychometric test [16]. The results showed that our model requires about 1,000 words to achieve an acceptable reliability ($\alpha \geq 0.8$) for 19 out of 35 Big 5 personality traits. This means that given a certain amount of data evidence (e.g., 1,000 words), our engine can infer a person's traits reliably.

Following many other studies in evaluating machine-inferred personality traits (e.g., see Refs [1, 35, 47, 51]), we also compared our inferred traits with personality traits measured using a traditional personality test. Like these works, we focused on reporting effect sizes, since such reporting facilitates the interpretation of the substantive as opposed to the statistical significance of a research result. It is considered a good practice when presenting empirical research findings [15].

To minimize inaccuracy or faking that may occur in a traditional personality test [108], we situated participants in a low-risk context. In particular, 475 student volunteers were recruited at a public university to take an IPIP-300 personality test⁴ and also chat with our AI interviewer Kaya for about 1 hour. We then computed Pearson linear correlation coefficients as effect sizes to quantitatively measure the strength of two datasets: our machine inferred personality traits and personality scores by IPIP-300. Our results showed that these coefficients (r) range from -0.088 to 0.294 for 35 Big 5 personality traits. Table 1 lists top-13 traits with coefficients equal to or greater than 0.1 because these are considered significant [15]. It is worth noting that our coefficients appear *stronger* than that in other studies (e.g., see Refs [1, 35]). For example, compared with our own previous work, the highest coefficient was 0.18 [35], while the current highest coefficient reached 0.29 (Table 1). Since both studies used the same IPIP-300 personality test, we attribute the differences to our respective personality inference engines. While our previous engine used LIWC word categories to predict Big 5 personality traits, our current engine uses much richer linguistic evidence to do so.

We also examined the negatively correlated traits, for example, *Vulnerability*, with the highest negative correlation. Due to social desirability biases, we are wondering how truthfully the students rated themselves on items, such as “I panic easily” or “I feel that I'm unable to deal with things,” even though in a low-stake context like ours. In contrast, our AI agent did not ask any such questions during an interview but automatically inferred the trait.

To test the validity of our inference model, we further gathered empirical evidence. First, the hiring managers were able to quickly find suitable candidates based on the ranking of their inferred traits (e.g., *Curiosity*). The follow-up in-person interview assessments also showed certain consistency between human-assessed and the inferred traits. For example, an interviewee's *Curiosity* observed by a human interviewer matched her inferred curiosity. Second, as described in Section 8,

⁴The test on the website <http://www.personalityassessor.com/ipip300/> is one of the most used personality tests to infer one's 35 Big 5 personality traits.

Table 1. Pearson Correlation Coefficients for the Top-13 Big 5 Personality Traits with Significant Effect Sizes between Machine-Inferred and IPIP-300 Measured

Intellectual Curiosity	0.294
Openness	0.265
Impulsiveness	0.239
Extroversion	0.22
Gregariousness	0.169
Cheerfulness	0.168
Liberalism	0.166
Friendliness	0.166
Cautiousness	0.166
Sympathy	0.16
Artistic Interest	0.115
Imagination	0.113
Conscientiousness	0.11

our analysis results present certain evidence that corroborates the validity of inferred traits. For example, we found that *Similar Personality* and *Extroversion* are significant factors impacting users' perceptions of Kaya. Since Kaya is designed to be an extrovert, a user with a similar personality should also be an extrovert, which is consistent with the inferred *Extroversion* of this user.

However, there is still much work needed to rigorously evaluate the validity of our personality inference model and such a task is non-trivial and often takes years of work [35]. This is because gathering ground-truth takes time and effort. For example, it takes time to validate whether a person is truly high on *self-efficacy* in the real world. One direction we are pursuing is to collect longitudinal data to examine how well the inferred traits predict users' real-world behavior, such as job performance or academic performance, especially in comparison with self-reported personality scores.

5 TOPIC-BASED CONVERSATION ENGINE

To power AI interviewers for practical uses, we have designed our conversation engine to meet three criteria. First, it must support diverse conversations and easy customizations of such conversations for a wide variety of interview situations, such as job interviews, patient inquiries, and customer surveys. Second, the engine must be easily extensible to accommodate new capabilities, e.g., extending an AI interviewer's ability to coach a user. Third, the engine must allow easy integration of third-party conversational technologies, e.g., Google Natural Language Processing (NLP) API.

With these design criteria in mind and inspired by cognitive architectures [53], we have developed a topic-based conversation engine that combines rule-based and algorithmic approaches to model and manage a conversation as a set of topics. In this approach, conversations can be easily customized to accommodate various interview situations through topic definitions. Moreover, conversational capabilities can be easily extended to incorporate third-party functions through topics.

5.1 Conversational Topics

We model a conversation as a set of topics, each of which defines a specific semantic theme of the conversation. Let us first define a conversational topic.

<pre>;; Define a topic that asks a user a question. ;; Here variable ?q holds a question to be asked. (deftopic ask-user-intro [?q] T1: [(chat-begin)] R1: [(ask-question ?q)] (handle-user-response ?q)) (deftopic handle-user-response [?q] T11: [(?input +)(short-response ?input)] R11: [(ask-for-more)] (handle-user-response ?q) ... T1N: [(?input +) (acceptable-response ?input)] R1N: [(agenda-topic-done)]) </pre> <p>(a)</p>	<pre>;; Define a topic that answers a user's inquiries (deftopic answer-job-inquiry T2: [when make decision] R2: ["We will make a decision..."] T3: [how many apply] R3: [(answer-num-candidates)]) </pre> <p>(b)</p>
<pre>;; Define a set of questions to be asked by a REP (question [user-intro-q { :type :open-ended :heading "Could you introduce yourself?" } ... job-preference-q { :type :single-choice :choices [{value 1 :text "customer service"} {value 2 :text "data analysis"} ...] }]) </pre> <p>(c)</p>	<pre>;; Define a REP's configuration for a virtual interview (config { :name "Kaya" :agenda [(ask-user-intro user-intro-q) (ask-fav-thing fav-movie) (ask-opinions opinion-q1) answer-job-inquiry ...] :fallback [:unordered movie sports] :exception [error-handling] :session-duration-max 60 ... }) </pre> <p>(d)</p>

Fig. 4. Various representations used in defining a REP and a virtual interview, including topics, questions, and configurations. Semantic units within a topic are highlighted in gray. Labels T1-T3 and R1-R3 are not part of the definition and added for the purpose of illustration.

Topic Representation. Semantically, a topic is similar to a discourse segment [13]. Figure 4 shows two example topics used in a job interview conversation: one soliciting initial information from a candidate (Figure 4(a)) and the other answering a candidate's questions regarding the position (Figure 4(b)). Each topic further consists of one or more *semantic units*, each of which handles a specific aspect of the topic.

The design of our semantic units is inspired by Artificial Intelligence Modeling Language (AIML), a language for programming chatbots [97], but improved for practical concerns (e.g., performance). Specifically, each semantic unit consists of a pair: a trigger (T) and a response (R). A *trigger* encodes one or more conditions under which a semantic unit may be activated during a conversation, while a *response* defines one or more corresponding agent actions. As shown in Figure 4(a), topic *ask-user-intro* has one semantic unit (T1-R1). Here, *trigger* T1 defines that the condition of asking a user for a self-introduction must be at the beginning of an interview. If this condition is met, the agent then asks the user to make a self-introduction (*response* R1). In Figure 4(b), topic *answer-user-inquiry* has two semantic units, one handling a user's inquiry on decision timeline (T2-R2) and the other on the number of candidates (T3-R3). These units define the respective conditions (triggers) and the corresponding agent responses.

For extensibility, both triggers and responses contain patterns of expressions and/or functions. For example, in R1, (*ask-question*) is a function and generates a text expression of a question. Figure 4(c) shows an example question, *user-intro-q*. Since a question definition often specifies only the heading of a question (see *user-intro-q*), function (*ask-question*) can transform a question heading into different expressions based on a conversational context. As described in Section 5.5, for example, it may phrase a question per Kaya's personality vs. that of Albert.

As shown in Figure 4(c), an agent may ask different types of questions during an interview. For example, question (*user-intro-q*) is an open-ended question to solicit a user's free-formed input, while (*job-preference-q*) is a choice-based question to ask a user's job preference. Choice-based questions are normally presented in a GUI form during an interview to simplify data processing.

Optionally, a semantic unit may be associated with a *sub-topic* that deepens a conversation on the topic. In Figure 4(a), (*handle-user-response*) is a sub-topic that handles a user's response to a question (?q). The definition of a sub-topic is the same as a regular topic, consisting of a set of semantic units. For example, in sub-topic (*handle-user-response*), semantic unit (T11-R11) states that if a user's answer (captured by *?input*) is too short, the agent should ask for more information and then recursively processes the user's new input. In addition, semantic unit (T1N-R1N) defines that if a user's response is acceptable, the agent should move on and mark this topic done by calling the function (*agenda-topic-done*). As shown in this example, sub-topics may be recursively invoked (R11) to support virtually arbitrary levels of conversations.

Topic Properties. To manage various topics during a conversation, each topic is further associated with a set of properties.

Importance. Just like in a human-human conversation, not all topics are equally important. Currently, we support three types of topics, ranked by importance. There are *agenda topics* that must be activated during a conversation at a certain point in time. As described below, in an interview, an AI interviewer activates an agenda topic to pose interview questions and collect needed information from a user. There are *fallback topics* that may or may not be activated during a conversation. These topics are mostly used to help make conversational transitions or fill the gaps during a conversation. For example, assume that an agent has finished asking all the interview questions and there is still time left. One of the fallback topics may be activated to let the agent continue with a user. *Error-handling topics* are only activated if errors or unexpected situations occur during a conversation.

Structure. Since topics can be defined hierarchically with nested sub-topics, sub-topics can only be activated if their associated parent topics are activated.

Temporal Order. Similar to a human-human conversation, certain topics (e.g., asking a user to introduce himself) are discussed before others (e.g., the user's opinion on mobile phones). A partial temporal order may be specified among defined topics to control the topic order. Topics may not be ordered at all to support an arbitrary flow among them.

5.2 Configuring an AI Interviewer by Topics

Using topics and sub-topics mentioned above, we model an AI interviewer in three steps. First, we specify a set of interview questions, like those defined in Figure 4(c). Second, we define a set of topics and subtopics to pose interview questions and handle user responses to the questions. For example, in Figure 4(a)–4(b), topic (*ask-user-intro*) asks and handles a user's self-introduction, while topic (*answer-job-inquiry*) handles users' questions about the position. Third, we configure an AI interviewer based on the defined topics and questions.

```

// activeTopics is a vector of all active topics. Initially, it includes sidetalk topics and error-handling topics
// specified in a configuration. See Figure 4(d)
// agendaTopics is a vector of topics specified on an agenda.

handle-agenda () {
1 while (agendaTopics is not empty) do
2 {
3   currentT=agendaTopics.getCurrent();
4   foreach topic[i] in currentT do
5   {
6     add topic[i] to activeTopics
  } } }

handle-active-topics () {
1 while activeTopics is not empty do
2 {
3   activeSUs = getMatched();
4   topSU = getTopMatch(activeSUs);
5   topSU.run();
  } }

```

(a)
(b)

Fig. 5. Pseudo code for our agenda-driven conversation management.

Figure 4(d) provides an example configuration. This configuration specifies four agenda topics, such as *ask-user-intro* and *answer-job-inquiry*, one activated after the other. In this configuration, the first three agenda topics have parameters, each of which represents an intended interview question. For example, topic *ask-user-intro* takes question *user-intro-q* (Figure 4(c)) as its parameter to ask a user to make a self-introduction. This parameterization enables these topics to be reused. In addition, there are two fallback topics, *movie* and *sports*, which may be activated in any order. This configuration also indicates the use of a special topic *error-handling* to handle various exceptions. A configuration can also define other properties of a REP, such as the name of an agent and the maximum duration allotted for an interview.

5.3 Running a REP by Agenda Topics and Topic Activation

Given a configured agent (Figure 4(d)), the conversation flow is managed through two parallel processes (Figure 5). The first process *handle-agenda* (Figure 5(a)) iterates through each defined agenda topic and activates one topic at a time, based on the specified temporal order (lines 1–5). Since a temporal order among these topics might be a partial order, it is possible that multiple agenda topics might be activated around the same time (lines 3–5). The activated agenda topics are then added to the *activeTopics* list to be processed (line 5).

To drive a conversation flow, the second process *handle-active-topics* (Figure 5(b)) examines all active topics and their associated properties, and decides which *semantic unit* under an active topic should be executed (lines 1–4). As described in Section 5.1, each topic contains one or more semantic units and each semantic unit contains a trigger that represents one or more conditions of a conversation. If a trigger is activated, the corresponding response defined in this semantic unit becomes a candidate for execution. For example, the state of a conversation may match trigger T1, while a user's input may match trigger T11 (Figure 4(a)). Accordingly, their corresponding responses (R1 and R11) are candidates for execution.

At a given time point of a conversation, however, one or more of the triggers may be matched. In such a case, the engine ranks them by their *structure*, *importance*, and *temporal order*, as mentioned in Section 5.1 (line 3, Figure 5(b)). Specifically, it ranks an agenda topic or its subtopic the highest by their temporal order, if any, followed by fallback and error-handling topics. This ranking allows a conversation to go deeper and guarantees that all agenda topics be discussed. By the ranking, the top-ranked semantic unit is then executed. The agent will then respond accordingly as defined by the semantic unit (e.g., R11). Depending on the definition of a response, the conversation may go deeper into a subtopic (e.g., R11 for topic *ask-user-intro*) or ends the current agenda topic (e.g., R1N). If the current agenda topic is ended, it is then removed from the agenda and the *activeTopics* list.

A conversation continues if there are still agenda topics to be processed (line 1, Figure 5(a)). Unless the conversation is explicitly terminated by one of the agenda topics, the conversation will still go on even after all agenda topics are done. This is because the *activeTopics* always include fallback topics, which can be activated by a user's input (Figure 5(b)), e.g., in a post-interview conversation.

5.4 Pattern-Based User Input Interpretation

Accurately interpreting a user's input is one of the biggest challenges in building a conversational system. Although our agent infers a user's traits based on observed linguistic cues (Formula 1) without the need to understand every user input precisely, it still must understand a user's input adequately to drive the interview process. To support the real-world use of an AI interviewer, we adopt a pattern-based approach to user input interpretation, since it has shown its effectiveness in many conversational systems [97, 100]. Each pattern is defined by a composition of tokens, strings, or regular expressions, which are recognized as states in a finite-state machine (FSM). We chose to use FSM to process users' natural language input for two reasons. First, we do not have interview data to train a language model. We must bootstrap our system by coding patterns of expressions (rules) to match potential user input. Second, FSM can capture diverse language patterns and prior work demonstrates its effectiveness in processing user input in conversation systems [67].

As detailed in Ref. [67], each pattern defined in a trigger (e.g., T2 and T3, Figure 4(b)) can be represented by an FSM. For generalization purposes, where a pattern can match with many input variants, our engine automatically lemmatizes tokens and inserts wildcards. For example, the pattern [how many apply] defined in trigger T3 (Figure 4(b)) can match a wide variety of user inquiring expressions regarding the number of candidates, such as "*how many candidates have applied for the job*" or "*how many applied for the position*."

It is worth pointing out that the main goal of our AI interviewers is to ask interview questions and obtain user responses. This is very different from many conversational systems of which the main goal is to perform a specific task (e.g., scheduling an appointment). In an interview context, if an agent fails to comprehend a user's response or answer a user's question, it can be guided to continue the conversation gracefully (e.g., skipping the current interview question).

Patterns are compiled into optimized FSMs by minimizing the number of states [67]. The FSM compiler itself is optimized by using several novel pattern rewriting techniques, such as recursively factoring out common prefixes and suffixes, retracting a large list of alternatives into a single special token, and eliminating duplicated patterns. As a result, hundreds of thousands of complex patterns can be compiled within seconds. Our current FSM runtime can process more than 100 million user input tokens per second on a mid-tier Macbook Pro Laptop.

5.5 Template-Based System Utterance Generation

Currently, we use a template-based approach to generate an agent's utterances during a conversation. To make an agent sound natural, an utterance template may include multiple alternatives. For example, a template that allows an agent to utter positive confirmations may include options, such as "*Great job*" and "*Well done*." Moreover, a template may consist of multiple segments. For example, here is a template that contains two segments, one explaining why the asking, while the other states the actual question: *ask-fav-movie-template* = <why-ask> <question>. Here <why-ask> may contain a set of alternatives:

```
{:option1 "I love magic so my favorite movie is..." :personality :cheerful}
{:option2 "Psychology studies show that..." :personality :serious}
```

Table 2. Two Virtual Personas and Their Behavior

	Kaya	Albert
Image Profile (Static)
Personality	Gregarious, Cheerful, Warm, Agreeable, Humorous; Like a friend	Reserved, Calm, Assertive, Rational, Careful; Like a counselor
Effective Inquiring	Affective strategy [36]; Positive politeness [12] Example: “You are so knowledgeable, would you mind telling me more about...”	Cognitive strategy [36]; Negative politeness [12] Example: “I’m sorry to keep you longer, would you mind telling me more about...”
Effective Influencing	Empathy, Comfort, Frankness, Cooperation, Agreement [11] Example: “I can certainly understand your point and agree with you that ...”	Reassurance, Commitment, Forgiveness [11] Example: “It is important to answer this question but no need to overthink...”
Small Talk	Personable [9] Example: “I love to chat with my online friends in my spare time, what do you do in your spare time?”	Minimal chitchatting Example: “Life is not just about work, what do you do in your spare time?”
Linguistic Style	Questions, suggestions, affective expressions [9, 36, 104] Examples: “Thank you so much for your input!” “Could you say a bit more?”	Assertions, projective statements, terse expressions [36, 72, 104] Examples: “Thanks.” “Please tell me more.”

As shown above, these alternatives are associated with a personality tag, which allows agent utterances to be customized to suit their personality. For example, **option 1** is used by cheerful Kaya to phrase the question, while **option 2** for serious Albert. Whenever suitable, utterance segments are tagged with different descriptors, such as a personality descriptor or style descriptor (e.g., affective vs. cognitive expressions, see more in Section 6.1). At runtime, the conversation engine selects suitable utterances by such descriptors to match the personality or style of an agent.

6 DESIGN OF A VIRTUAL INTERVIEWER

To answer our research questions, we developed two AI interviewers, each with a distinct personality that is associated with effective human interviewers [45, 104].

6.1 Two Virtual Personas

Existing studies suggest two types of effective human interviewers: one with a warm, cheerful personality, effective at building rapport with interviewees, and the other with a serious, assertive personality, effective at requesting cooperation from interviewees [45, 104]. We thus designed and built two virtual personas. Table 2 summarizes the two AI interviewers and their matched behavior.

Appearance. Since people often associate one's appearance with a particular personality [48, 77, 101], we created two static profile images, each representing a distinct persona. In particular, Kaya takes the form of a cheerful female, while Albert is a serious-looking male (Table 2). Our gender assignment is based on psychology studies, which show women in general are warmer and more sensitive in their personality, while men are more reserved and dominant [32, 99].

Conversational Behavior. In addition to appearance, an agent's conversational behavior exhibits its personality [11]. Specifically, we unveil the personality of an AI interviewer through four types of conversational behavior associated with an interview task.

Effective Inquiring. Clinical studies show that interviewers can use different techniques to effectively draw out information from interviewees. For example, an *affective strategy* helps an interviewer establish rapport with an interviewee, which then allows the interviewee to open up and act authentically [36]. On the other hand, *negative politeness* may be used to preserve the emotional distance between an interviewer and an interviewee, minimizing imposition on the interviewee [12]. Since the use of particular strategies or techniques reveals an interviewer's personality, we match our virtual personas with different inquiring strategies or techniques. Specifically, aligning with her warm and cheerful personality, Kaya uses an affective strategy and positive politeness in a conversation, while Albert uses a cognitive strategy and negative politeness to reveal his reserved and serious personality. Table 2 lists two contrasting examples that show Kaya's affective expression with positive politeness versus Albert's negative politeness.

Effective Influencing. Besides eliciting information, an interviewer may also want to influence interviewees (e.g., encouraging them to think outside the box) [72]. To achieve this goal, we introduce influence operators [11] to help virtual interviewers effectively guide interviewees to perform their best. To exhibit an interviewer's personality, we assign different influence operators: Kaya uses empathetic and cooperative language, while Albert's language is reassuring and forgiving. See expression examples used by Kaya and Albert, respectively, in Table 2.

Small Talk. While small talk helps make a conversation more engaging [11], it may also block honest interaction [37]. To match their personalities, Kaya uses small talk to build rapport and show her friendly side, while Albert rarely chitchats to reflect his serious persona. In our current implementation, small talk is incorporated as part of an interview question or a response to a user's input. See the contrasting sample expressions used by Kaya and Albert in Table 2.

Linguistic Style. Since one's linguistic style unveils one's personality [74, 81], we incorporate matching linguistic cues (e.g., words, phrases, and punctuations) into our AI interviewers' utterances. These linguistic cues are selected based on previous studies [47, 49, 104] and our own findings that establish the relationships between various linguistic cues and personality traits (Formula 1). For example, Kaya frequently uses first-person, affective expressions ("I love romantic movies as they make me cry."), emoticons, and exclamations. In contrast, Albert uses third-person declarative, projective statements ("It makes sense.") or abbreviated replies ("Understood."). Moreover, Kaya uses questions and suggestions to show her friendliness, while Albert uses assertions and demands to display his seriousness.

As mentioned in Section 5.5, we use a template-based approach to agent utterance generation. We have implemented a set of templates to cover common conversational acts, such as *greeting* and *acknowledging* [83]. To reflect the personality of an AI interviewer, we have created alternative expressions within a template based on various strategies mentioned above and associate each with a corresponding personality descriptor (see Section 5.5). For example, an *acknowledging* template contains *thank-you* expressions corresponding to different personalities (e.g., examples in the last

row of Table 2). During runtime, our system selects the ones that match the given personality of an AI interviewer.

6.2 Interview Questions

To create a customized AI interviewer, typically an organization first provides us with a set of interview questions pertinent to their context, such as inquiring about a job candidate's interests and skills related to a position in a job interview or asking a student's experience with a class in an educational setting. We put in additional remarks and questions to: (1) make the interview flow more naturally, (2) measure candidates' willingness to confide in and listen to a virtual interviewer, and (3) solicit candidates' perception and feedback of the AI interviewer. Although each field deployment consisted of very different interview questions, each interview flow is similar with five parts shown below.

- (I) *Introduction*. The first part includes a set of open-ended questions, helping an agent build rapport with a user. It typically starts by asking the user to make a self-introduction, and then chats with the user to collect additional information, such as the user's hobbies, favorite movie, and role model.
- (II) *Impression Management Questionnaire*. The second part includes a 20-item impression management (IM) scale [34] to measure how much one inflates his/her own image by exaggerating desirable responses especially in a high-stakes situation like a job interview. We also want to use this measure to examine an interviewee's willingness to be true to him/herself during an interview.
- (III) *Opinion Questions*. This part normally includes a set of open-ended questions that solicited a user's input or opinions on various subjects related to a specific interview situation. Some of these questions were controversial in nature, aiming at measuring a user's willingness to confide in and listen to an agent (see more below).
- (IV) *About You*. This part includes a set of questions, asking a user more information about herself, such as her job preferences, skills and interests, and strengths and weaknesses. The agent often shares and discusses its analysis of a user, such as her personality traits, top strength, and top weakness. The agent may also suggest certain actions, which were meant to measure users' willingness to listen to the agent (see more below).
- (V) *Post Interview Survey*. This last part collects demographics from users and their rating of the REP on several aspects, from personality to usability (see more below).

Measuring Users' Willingness to Confide in an Agent. Drawing out authentic information from interviewees is paramount to any successful interview, since such information helps both interviewers and interviewees better assess real-life situations (e.g., the fitness of a candidate for a particular job) [72]. However, due to social desirability biases—especially in high-stakes situations like job interviews—interviewees may not be completely truthful [73, 106]. While studies show that users are more willing to confide in machines [59], we want to further understand how the personality of an agent and the personality of a person affect one's willingness to confide in the agent, especially in high-stakes situations.

To answer our first research question, we assess one's willingness to confide in an agent by their willingness to share sensitive information with others, such as a hiring manager. In particular, we measure such willingness through two sets of measures in an interview: the Impression Management (IM) scale, and asking the users to take actions to reveal and share sensitive information about themselves or their opinions on controversial topics with a third party, e.g., a hiring manager.

Table 3. Example Actions that Measure One's Willingness to Confide in and Listen to a REP

$A = \text{rating} \times \text{sAction}$	Click-through ($i = 1, N$) $C_i = \begin{cases} 0, & \text{URL not clicked} \\ 1, & \text{URL clicked} \end{cases}$
Share Weakness: $\text{rating} = \begin{cases} 3, & \text{disagree} \\ 2, & \text{not sure} \\ 1, & \text{agree} \end{cases}$ $\text{sAction} = \begin{cases} 0, & \text{don't share} \\ 1, & \text{share mine} \\ 2, & \text{share REP sugg.} \end{cases}$	
$B_j = CF_j \times \text{sAction}_j$	$S_j = \text{rating}_j \times \text{act}_j$
Share Opinions ($j = 1, K$) $CF_j = \begin{cases} 1, & \text{High} \\ 2, & \text{Med} \\ 3, & \text{Low} \end{cases}$ $\text{sAction}_j = \begin{cases} 0, & \text{don't share} \\ 1, & \text{share} \end{cases}$	Share Action ($j = 1, M$) $\text{rating}_j = \begin{cases} 3, & \text{disagree} \\ 2, & \text{not sure} \\ 1, & \text{agree} \end{cases}$ $\text{act}_j = \begin{cases} 1, & \text{share REP's} \\ 0, & \text{share mine} \end{cases}$
(a) Willingness to confide: $WC = A + \sum_{j=1}^K B_j$	(b) Willingness to listen: $WL = \sum_{i=1}^N C_i + \sum_{j=1}^M S_j$

The IM scale [73], also known as a “lie” or social desirability scale, is designed to detect how users consciously favor themselves to impress others. A user rates oneself using a 7-point Likert scale on each of 20 statements, such as “*I have never driven over a speed limit.*” If a user rates oneself on the scale of 6 or 7 on this statement to indicate that s/he has never driven over speed limit, it is most likely that s/he consciously inflates oneself since the statement describes an extreme situation. In this case, s/he receives one point. After going through twenty statements, a user receives an accumulated total score. The higher the score, the more inflation there is, and thus, the less willing one is to confide in another with honest ratings [73].

In addition to the IM scale, we designed user actions to measure one's willingness to confide in an agent. In our interview context, each action is related to sharing certain sensitive information or an opinion on a controversial subject with a third party who can influence an interviewee's life (e.g., a hiring manager or an academic advisor).

Table 3(a) provides two example actions used in our field deployments. The first one is to share a user's biggest weakness. The agent asks a user about her biggest weakness before showing its inferred top weakness for the user. The agent then asks the user to rate the inferred weakness by a 3-point *rating* (Table 3(a)). Heuristically, this rating measures the “*sensitivity*” of the information to be shared. If a user disagrees with the system-suggested weakness, this weakness information might be more sensitive and a user might feel more risky to share such a weakness with someone who has influences. As part of the action, the user must choose one of three actions (*sAction* in Table 3(a)). The user's rating and choice of action are then used to compute the user's willingness to confide in an agent (*A* in Table 3(a)). By this measure, a user is most willing to confide in the agent if she shares her weakness suggested by the agent, even if she *disagrees* with its analysis.

Another example of a sharing action is to solicit a user's opinion on a controversial subject (e.g., “*What important truth do very few people agree with you on?*”). In our field deployments, our clients wanted to use such questions to assess an interviewee's soft skills from sources like Ref. [89]. On a controversial subject, the agent first asks a user's opinion and then probes further by eliciting a rationale or offering counter viewpoints. The agent then asks the user's confidence in her opinion (*CF* in Table 3(a)) before asking her action (*sAction* in Table 3(a)). Per the user's reported confidence and choice of action, the willingness to confide in is then computed (Table 3(a)).

As seen from the example actions above, we designed user actions to measure one's willingness to confide at multiple scales instead of as a binary value (sharing vs. no sharing). Moreover, we design the actions to measure the “*sensitivity*” of the information to be shared. The less confidence a user has in the information, the more sensitive the information is. By our measure, a user is most willing to confide in an agent if she shares the most “*sensitive*” information with a third party.

Our rationale is that people always want to show their best qualities, especially in a high-stakes situation like a job interview. When they have an option not to share something they are uncertain of, they could hold back without risking exposure of their potential weaknesses unless they are willing to confide in an agent despite the risks.

The sharing actions described above are just examples used in our field deployments. Such actions may be adapted for a specific interview situation by the same design principles mentioned above. For example, one of our real-world deployments is to interview university students to find out how their personality may be related to their academic interests and performance. Instead of asking a student's willingness to share his/her top weakness, the question is revised to ask the candidate's willingness to share with his/her academic advisor who would give him/her the worst reference. Although the questions may be different, the goal is the same: it is to measure a user's willingness to confide in by sharing sensitive information. When designing such actions, we should also carefully balance between being able to distinguish a user's behavior and the required user cognitive effort, especially when an interview is involved and long. This is also why we use a simple 3-point scale to distinguish a user's behavior (Table 3).

Measuring Users' Willingness to Listen to an Agent. In addition to eliciting authentic responses from interviewees, an interviewer should also effectively guide the interviewees and help them reveal their best qualities (e.g., demonstrating *thoughtfulness*). However, users may not always be willing to listen to machines [77, 78]. To answer our second research question and find out what might affect users' willingness to listen to an agent, we designed a set of user actions using the similar principles described above to measure one's willingness to listen to an agent. Table 3(b) shows two sample actions and their measures.

Again, such user actions may be adapted for different interview situations. Here, we give several example actions used in our field deployments to demonstrate how we measure one's willingness to listen. In one such action, *willingness to listen* is measured by a user's click-through after an agent asks the user to read online articles related to the current interview topic by clicking on the articles' URLs.

Another type of action is related to persuading a user to share a piece of agent-suggested information. Similar to the classic Desert Survival Problem that is used to test the power of persuasion [78], this type of action that tests one's willingness to listen has four steps. The agent first asks a user's input on a specific subject (e.g., her top skill) before offering a suggestion, which is often different from the user's input. The agent then asks the user to rate its suggestion (*rating* in Table 3(b)) before persuading her to share its suggestion, as opposed to the self-reported one, with a third party, who can often influence the user's real life. In such an action, *willingness to listen* is measured by how strongly a user disagrees with an agent's suggestion and the user's final choice of action.

Measuring Users' Perceptions of an Agent. Based on previous studies on anthropomorphic interfaces [54, 77, 86, 87], we have designed four sets of Likert-scale questions to measure users' perceptions of an AI interviewer. The first set asks users' perceived trustworthiness of an agent, since such perceptions may be directly related to their willingness to confide in and listen to the agent. To validate the effectiveness of our personality manipulation (i.e., whether users have perceived an agent's given personality), the second set asks users' perceptions of an agent's personality from multiple aspects, such as perceived characteristics (e.g., how *cheerful* or *reserved* an agent is), role similarity (i.e., how an agent is similar to a friend or a counselor), and personality similarity (i.e., how similar a user's personality is to an agent's). The third set asks a user to rate the quality of an agent from several aspects (e.g., likeability and insightfulness). The last set solicits a user's feedback on an agent's usability (i.e., how enjoyable it is) [78].

7 FIELD DEPLOYMENTS

As described in Section 6, we have designed two distinct interviewer personas, warm and cheerful Kaya and serious and assertive Albert, each of which aims at achieving a particular interview goal that we study. Specifically, based on previous findings [45, 106], Kaya would be better at establishing rapport with users and encourages users to confide in this agent, while Albert would be better at demanding cooperation and encourages users to listen to this agent. In addition, we have carefully designed a set of interactions that users must perform during an interview so that we can measure their behavior with each AI interviewer persona.

Kaya and Albert were used in a number of real-world deployments. For the purpose of answering our research questions, we present four relevant deployments, involving a total of 1,280 interviewees. Among these four deployments, two of them were conducted in high-stakes situations for real job interviews, while the other two were done in lower-stakes situations for educational purposes. Across all four deployments, each user spent well over an hour with an agent and contributed about 1,600 words on average.

7.1 Interviewing Entry-Level Job Applicants (Deployment I)

In this field deployment, we deployed both Kaya and Albert to aid a firm in hiring an entry-level associate. The firm received about 800 applications and notified all the applicants to expect an email invitation of a virtual interview opportunity and encouraged them to use this opportunity to showcase their unique characteristics. Half of the candidates were invited to interview with Kaya, and the other half to interview with Albert. Each email invitation contained an individualized URL to start the interview at our website. Each interview contained five parts (Section 6.2) and questions in Part I (Introduction) and Part III (Opinion Questions) were customized by the hiring manager to suit the hiring context.

The process lasted about three weeks, during which 157 applicants completed their virtual interviews with Kaya, 159 completed theirs with Albert, and 59 people started but did not finish their interviews. Among the 157 users with Kaya, there were 130 (83%) male and 27 female (17%); among the 159 users with Albert, there were 125 (79%) male and 33 (21%) female. Users' ages ranged from 18 to over 50, and 90% of them were between 18 to 35 years old. On average, each user spent about 75 minutes interviewing with Kaya or Albert.

7.2 Interviewing Summer Internship Applicants (Deployment II)

Similar to Deployment I, this field deployment helped a high-tech firm interview their summer internship applicants. Because all their applicants are university students, the firm believed that Kaya in a younger, warmer persona would better engage with the student population than Albert would and decided to use only Kaya for all the virtual interviews. In this case, each internship applicant first submitted an application via email. S/he then received an automatic email response that contains a URL where the applicant could sign up and complete an interview with Kaya before a deadline.

Similar to Deployment I, each interview contained five parts (Section 6.2) with questions customized to suit the hiring context. By the deadline, a total of 454 students signed up, 307 students started their virtual interview, and 290 of them completed theirs. Among those who completed, ages ranged between 18–35, 89 (31%) of them were female, 201 (69%) were male, and 186 (64%) were international students. On average, each applicant spent about 1.5 hours with Kaya.

7.3 Interviewing Students for Team Experience and Preferences (Deployment III)

Unlike the two deployments described above, this deployment was conducted by a professor in a big public university to: (1) gather information from the students in his class about their team

project experience and preferences, and (2) infer the students' personality traits. The purpose was to help students find better teammates in class projects. For the same reason as described above, Kaya was chosen as the AI interviewer for the student population. Like the interviews in Deployments I & II, each interview also contained five parts (Section 6.2) but with questions customized to suit the context. For example, in Part III (Opinions), Kaya asked each student a series of team-related questions designed by the professor, e.g., "*What have you learned from the past team experience?*"

There were 201 students in the class, all of whom completed an interview with Kaya as one of their class assignments. Since eight (8) students did not submit their consent form, we excluded their data from our analysis. Among the 193 students, all of them between 18–25 years old, 145 (75%) were male, 48 (25%) were female, and 70 (36%) were international students. On average, each user spent about 90 minutes interviewing with Kaya.

7.4 Interviewing Students for Assessing Student Personality Traits (Deployment IV)

The fourth deployment was also conducted in another big public university. It aimed at using interviewers to infer students' personality traits and investigate how the inferred traits may be related to the students' academic performance. Again, Kaya was chosen as the interviewer for the student population. Likewise, Kaya conducted each interview in five parts (Section 6.2) with customized questions. For example, for opinion questions (Part III), Kaya asked each student a series of school-related questions, such as "*What's your favorite class?*"

In this deployment, each student volunteer received extra course credits upon the completion of an interview with Kaya. A total of 473 students completed an interview with Kaya, all between 18–25 years old. Among the students, 351 (74%) were female, 122 (26%) were male, and 20 (6%) were international students. On average, each student spent around an hour with Kaya.

8 RESULTS

We first present an overview of the results and then present our analysis answering our three research questions.

8.1 Overview of Results

In Deployment 1, after each interview, we asked users to rate their perception of the REP's personality traits (on a scale of 1–5). The top three rated personality traits for Kaya were *calm* (mean = 3.89, std = 1.04), *cheerful* (mean = 3.875, std = 1.01), and *warm* (mean = 3.82, std = 1.09), and for Albert, *calm* (mean = 4.15, std = 0.88), *rational* (mean = 3.96, std = 0.93), and *assertive* (mean = 3.58, std = 0.94). Except for the highly rated *calm* for Kaya, users' perceptions matched with the intended personality of Kaya and Albert (Table 2). We attributed Kaya's perceived *calmness* to her role as a job interviewer. In Deployments II, III, and IV, instead of rating Kaya's personality traits, each interviewee was asked to describe Kaya's personality in three keywords. The idea was to learn what descriptors users would use.

We collected the keywords and computed their frequencies. Figure 6 shows the top-10 most frequently cited keywords from the three deployments. As shown, the interviewees' descriptions about Kaya across deployments were consistent: the most mentioned words were *friendly*, *kind*, and *nice*, which matched with our persona design of Kaya. In Deployment III, *robotic* was also used frequently to describe Kaya. As shown below in Table 4, users in this deployment also rated Kaya the lowest across almost all the dimensions. We were wondering whether the user reaction had something to do with the interview context: in this deployment, the students were required to chat with Kaya by their professor as one of their homework assignments, while users in other deployments opted in to chat with Kaya.

friendly (84), polite (41), kind (23), curious (19), nice (18), helpful (17), inquisitive (17), patient (16), smart (14), direct (12), intelligent (12)	friendly (43), robotic (24), kind (18), nice (16), polite (16), repetitive (11), patient (9), cheerful (8), straightforward (8), curious (7),	nice (85), friendly (66), kind (61), curious (54), interested (35), helpful (31), understanding (30), caring (28), smart (28), inquisitive (27)
(a) Deployment II (290 users)	(b) Deployment III (193 users)	(c) Deployment IV (473 users)

Fig. 6. Top-10 words most frequently used by users to describe Kaya's personality.

Table 4. Results of User Ratings on REP (* $p < .05$)

User Perceptions of REP	Deployment I (Job Interview)		Deployment II (Job Interview)	Deployment III (University)	Deployment IV (University)
	Kaya	Albert	Kaya	Kaya	Kaya
Like a friend*	mean = 2.7, sd = 1.2	mean = 2.4, sd = 1.2	N/A	mean = 2.5, sd = 1.3	mean = 2.6, sd = 1.3
Like a counselor	mean = 3.2, sd = 1.2	mean = 3.2, sd = 1.3	N/A	mean = 3.2, sd = 3.3	mean = 3.6, sd = 1.3
Helpful	mean = 3.6, sd = 1.1	mean = 3.6, sd = 1.0	mean = 4.0, sd = 0.99	mean = 3.1, sd = 1.3	mean = 3.5, sd = 1.2
Likable	mean = 3.5, sd = 1.1	mean = 3.3, sd = 1.0	mean = 3.97, sd = 1.0	mean = 3.2, sd = 1.3	mean = 3.6, sd = 1.1
Insightful	mean = 3.5, sd = 1.2	mean = 3.6, sd = 0.99	mean = 3.7, sd = 1.2	mean = 3.1, sd = 1.4	mean = 3.6, sd = 1.1
Interesting	mean = 3.8, sd = 1.2	mean = 3.9, sd = 1.1	mean = 3.97, sd = 1.2	mean = 2.8, sd = 1.3	mean = 3.5, sd = 1.2
Enjoyable	mean = 3.6, sd = 1.2	mean = 3.4, sd = 1.2	mean = 3.8, sd = 1.2	mean = 2.5, sd = 1.3	mean = 3.2, sd = 1.2
Trust	mean = 3.7, std = 0.99	mean = 3.8, sd = 0.8	mean = 4.0, sd = 0.8	mean = 3.5, sd = 0.99	mean = 3.8, sd = 0.83

The Highest Rating in Each Rating (row) Is Highlighted in Yellow, While the Second Highest Rating Is Highlighted in Gray.

Table 4 summarizes various user ratings for all the four deployments. In Deployment I, where both Albert and Kaya were used, an independent samples t -test showed no significant differences in most of user ratings between the two agents except that Kaya was rated to be significantly more *like a friend* than Albert ($t(307) = 2.54, p < .01$). Each agent was also rated on three trust measures (*integrity*, *competence*, and *benevolence*) [64], which were combined additively to create an index measure called *Trust*. As shown in Table 4 (the last row), users trusted Albert more than Kaya, and an independent samples t -test showed a trend: $t(314) = 1.88, p < .06$.

Across all four deployments, it is interesting to observe that job interviewees rated their impression of an agent higher than university students across almost all the dimensions. This might be because they considered the AI interview part of their application process and were more engaged with and positive about the agent. Moreover, summer internship applicants in Deployment II were more positive about Kaya than the job applicants in Deployment I. This might be because most of the intern candidates were Computer Science majors and were more familiar with and valued more the concept of AI agents.

Furthermore, in Deployments II, III, and IV, users were also asked to indicate their communication preference with an agent. Among the 964 users who completed an interview in these three deployments, 800 (83%) preferred the current text-based chat, while 164 (17%) would have preferred a voice-based conversation with an agent.

R1: Effects of Agent Personality on Users' Willingness to Confide In

Our first research question is to examine the effects of agent personality on users' willingness to confide in an agent. Since only Deployment I used two different agents, our analysis below analyzed the data from Deployment I.

As mentioned earlier, we measured one's willingness to confide in an agent from two aspects: the IM Scale [73] and a set of actions measured by an index variable *Willingness to Confide* (Table 3(a)).

By the IM Scale, users were less likely to want to impress Albert (mean = 12.68, sd = 3.48) than Kaya (mean = 12.93, sd = 3.44), although this difference was not significant. Using the IM scale as a dependent variable, we ran a General Linear Model (GLM) with three independent variables⁵: *Trust* (user perceived trustworthiness of an agent), *Similar Personality* (how similar users felt the agent's personality was to their own), and *Agent* (distinguishing Kaya from Albert), controlling for user's *Gender*. The result showed *Trust* was the only significant factor: $F(1,302) = 5.65$, $p < .02$, $Coeff = .59$. This implies that the more the users trust an agent, the more likely they want to consciously impress the agent. This finding seems consistent with the theory that people apply the same social rules to computers as they apply to humans [78]. In other words, people often want to impress someone whom they trust (e.g., parents and spouses).

With *Willingness to Confide* as a dependent variable, we ran a GLM with the same three independent variables: *Trust*, *Similar Personality*, and *Agent*, also controlling for *Gender*. It showed that *Agent* was significant: $F(1, 302) = 5.38$, $p < .02$, $Coeff = .46$. This implies that users are more willing to confide in Albert and they are more willing to do so if they trust the agent more. *Gender* was insignificant.

Our results show that users are more honest with (per the IM score) and more willing to confide in Albert. Initially, we hypothesized that users would be more willing to confide in an agent like Kaya who should be better at building rapport based on previous work (e.g., see Ref. [9]). However, our results are the *opposite*.

R2: Effects of Agent Personality on Users' Willingness to Listen

Our second research question is to examine the effects of an agent personality on users' willingness to listen to an agent. Since only Deployment I used two agents, the analysis used the data only from Deployment I.

Using an additive index of users' actions, *Willingness to Listen* (Table 3(b)) as the dependent variable, we ran a GLM over the same independent variables: *Trust*, *Similar Personality*, and *Agent*, controlling for *Gender*. It showed that *Agent* was significant: $F(1, 302) = 4.15$, $p < .04$, $Coeff = .39$. This result showed that users are more willing to listen to Albert, an agent with a reserved, assertive personality. However, since *Similar Personality* was not significant ($p < .09$), users may not necessarily be willing to listen to an agent who possesses a similar personality. We thus are unable to confirm previous findings [78, 87].

R3: Effects of Interview Context and User Individual Differences on User Behavior

To answer our third research question, we examined whether and how the context of a virtual interview and users' individual differences, such as gender and personality traits inferred from the interview, influence their perception of an agent as well as their behavior with an agent. We

⁵We included these three variables based on our research questions (e.g., including *Agent* to directly test the effect of agent personality on a dependent variable) and our preliminary analyses that examined the effect of each independent variable (e.g., users' various ratings of an agent) on a dependent variable. *Trust* and *Similar Personality* stood out as the most influential variables in such analyses. Moreover, previous literatures back up their potential effects on users' behavior [29, 64, 68].

analyzed two datasets: (1) the Kaya dataset that consisted of Kaya interviews in all the four deployments, and (2) the Albert dataset in Deployment I. As shown below, we examined the effects of a set of independent variables on user perception and behavior with Kaya and Albert, respectively.

Interview Context (*high-stakes* vs. *low-stakes*). Numerous studies show that context (e.g., leisure vs. work) affects users' perception of virtual agents (e.g., see Refs [68, 86]). In addition, research shows that interview context influences a user's willingness to confide in or volunteer truthful information to a computer (e.g., see Refs [27, 73, 108]). For example, a person is less likely to inflate oneself or lie in a low-stakes context [73]. Since our goal is to create effective agents who can elicit authentic information from users in any context, we wanted to examine how an interview context might influence users' perception of and behavior with an agent.

Our four deployments happened to be situated in two different contexts: real-world job interviews and surveys with university students. The first context is a high-stakes context as a job was at stake based on the interview. The second context is low-stakes, where the students had little at stake to impress or to deceive the agent in order to increase their stature. We thus examined how these two contexts affect a user's perceptions of and behavior with an agent.

Individual Differences. As shown in existing work, individual differences, such as gender and cultural background, affect one's perceptions of and behavior with virtual agents. In our analysis, we used four measures of individual differences.

Similar Personality. Previous research suggests that people who view an agent's personality similar to theirs should rate the agent higher [69]. We wanted to know whether a higher score on this measure would indicate a higher trust, higher agent quality, or higher enjoyment rating of an agent.

Culture. Cultures influence people's perceptions of and behavior with virtual agents [26]. The interviewees involved in our field deployments had different cultural backgrounds. For example, 64% of job applicants for the summer internship positions were international students. We were curious about whether their background would impact their perceptions of or behavior with a REP. For this analysis, one's cultural background was coarsely determined by one's native language since each user was asked of his/her native language during their interview. Based on the data collected, each user was assigned with one of two "culture" labels: "western culture" (associated with native languages such as English and Spanish) and "eastern culture" (associated with native languages such as Chinese or Indian).

Gender. We controlled for *Gender* since previous research indicates that females and males may perceive and respond to an agent differently [86].

Inferred Personality Traits. We expected that users' personality traits, inferred from their text during an interview, would affect their perceptions of and behavior with an agent. We performed a factor analysis on the 35 inferred Big 5 personality traits using a Varimax rotation with a Kaiser normalization for the Kaya and Albert datasets, respectively. For the Kaya dataset, a scree plot⁶ revealed that seven factors should be used, accounting for 51.56% of the variance (Table 5(a)). For Albert, a scree plot revealed seven factors accounting for 51.83% of the variance (Table 5(b)).

Effect of Interview Context and Individual Differences on User Perception of a REP. A user's perception of an agent was measured by three variables: (a) perceived trust, (b) perceived qualities of an agent, and (c) enjoyment of chatting with an agent.

⁶ A scree plot is used to determine the number of factors to select based on a visual analysis of when the curve flattens.

Table 5. Personality Traits Loaded onto Separate Factors

Factor 1: Extroversion ← Cheerfulness; Impulsiveness Factor 2: Openness ← [Liberalism, Intellectual Curiosity, Imagination] Factor 3: Neuroticism ← [Anxiety, Vulnerability] Factor 4: Conscientiousness ← Dutifulness Factor 5: Trust Factor 6: Cooperation; Sympathy Factor 7: Excitement-Seeking	Factor 1: Emotionality; Anxiety; Vulnerability Factor 2: Neuroticism ← [Hostility, Depression] Factor 3: Agreeableness ← Altruism Factor 4: Openness ← [Liberalism, Intellectual Curiosity, Imagination] Factor 5: (-)Adventurousness; (-) Self-efficacy; Immoderation Factor 6: Conscientiousness ← [Achievement-Striving, Self-Discipline] Factor 7: Extroversion ← [Excitement-Seeking, Assertiveness]
(a) Trait factors identified for Kaya dataset	(b) Trait factors identified for Albert dataset

Arrows Indicate Factor and Facets Relationship. The First Trait Mentioned in Each Factor Is Used as a Reference to Characterize the Traits that Loaded onto that Factor.

- *Perceived Trust*. As mentioned before, the *Trust* index measures users' perceived trust in an agent by additively combining their 5-point Likert scale ratings on an agent from three aspects: *benevolence*, *integrity*, and *competence*.
- *Perceived Agent Quality*. Each interviewee was asked to rate an agent on a 5-point Likert scale from four aspects: how *insightful*, *helpful*, *likeable*, and *interesting* the agent was perceived. Since all the four metrics were significantly correlated, we created an additive index of these measures called *Agent Quality Rating* to measure a user's perceived agent quality.
- *Perceived Enjoyment*. Each interviewee was asked to rate how much they enjoyed chatting with an agent on a 5-point Likert-scale. We use this *Enjoyment Rating* to measure one's perceived enjoyment.

We examined the effects of interview context and individual differences on each of these three variables. To do so, each of these three variables was entered in a stepwise regression model⁷ as a dependent variable with interview context and individual difference measures described above as independent variables. We used forward selection stepwise regression to find the best predictive model. All the results are summarized in Table 6. The independent variables entered for the Kaya model are different from those entered for the Albert model. This is because the Albert dataset was collected under only one context (Deployment I for job interview) and did not have users' cultural background information.

On modeling users' perceived trust for Kaya, our stepwise regression results found that *Similar Personality* and personality factors 1 (*Extroversion* and *Cheerfulness*) and 7 (*Excitement-Seeking*) significantly influenced one's perceived trust in Kaya. The more similar the personality, the more extroverted and cheerful (like Kaya), and the *less* one seeks excitement, the more trust one has in Kaya. Note that our users' personality traits are automatically inferred. This finding also validated our trait inference: users who have a personality similar to Kaya are indeed extroverted and cheerful as expected.

On the other hand, on modeling users' perceived trust in Albert, we found that *Similar Personality* and personality factor 4 were significant. The *less* open (more conservative) they were, the more they trusted Albert. This is very plausible due to Albert's personality since users might perceive

⁷Stepwise regression is a technique that enters and removes variables, particularly if there is a large set of independent variables like we used, to find the best predictive model [18]. In forward selection, variables are entered into the model one-by-one, based on choosing variables that improve the fit of the model the most, until no new variables provide a statistically significant improvement in the fit of the model.

Table 6. Interview Context and Individual Differences Predicting User Perceptions of a REP

User Perceptions of a REP	REP	Significant Factor	B(SE)	t	F (df)
Perceived Trust (benevolence, integrity, competence)	Kaya	Constant	2.93(.06)	52.74***	$F(3, 1097) = 99.33$ adj. $R^2 = .21$ $p < .0001$
		Similar Personality	.36(.02)	16.79***	
		Factor 1 (Extroversion)	.06(.02)	2.46*	
		Factor 7 (Excitement-Seeking)	-.08(.02)	-3.54***	
	Albert	Constant	3.26(.14)	23.98***	$F(2, 150) = 16.85$, adj. $R^2 = .17$ $p < .001$
		Similar Personality	.29(.06)	5.16	
		Factor 2 (Openness)	-.18(.06)	-3.03	
Perceived REP Quality (helpful, insightful, likeable, interesting)	Kaya	Constant	8.93(.23)	38.84***	$F(1, 1094) = 664.59$ adj. $R^2 = .38$ $p < .0001$
		Similar Personality	2.27(.09)	25.78***	
	Albert	Constant	10.90(.52)	20.87***	$F(1, 151) = 57.03$, adj. $R^2 = .27$ $p < .0001$
		Similar Personality	1.63(.22)	7.55***	
Perceived Enjoyment (enjoy chatting w/ a REP)	Kaya	Constant	1.71(.07)	23.13***	$F(2, 1094) = 298.08$ adj. $R^2 = .34$ $p < .0001$
		Similar Personality	.62(.03)	21.56***	
		Interview Context	.27(.07)	4.18***	
	Albert	Constant	2.27(.19)	11.74***	$F(1, 151) = 44.96$, adj. $R^2 = .22$ $p < .0001$
		Similar Personality	.54(.08)	6.71***	

For Modeling the Kaya Dataset, the Independent Variables Are Interview Context, Culture, Gender, Similar Personality, and Derived Personality Trait Factors Shown in Table 5(a). For Modeling the Albert Dataset, the Independent Variables Are Gender, Similar Personality, and Derived Personality Trait Factors Shown in Table 5(b).

Albert as being conservative, and that they assess their personalities (conservative) as being more similar to Albert. In this sense, our trait inference engine again correctly identified those with a more conservative characteristic.

On modeling users' perceived agent quality, for both Kaya and Albert, we found that only *Similar Personality* was a significant factor. In other words, users who felt having a personality similar to an agent rated the agent higher on its quality.

On modeling users' perceived enjoyment, again for both Kaya and Albert, the results showed that *Similar Personality* was significant. In other words, if one has a personality similar to an agent, the person would enjoy chatting with the agent more. For Kaya, *Interview Context* was also significant. Users might have invested more in interacting with an agent in a high-stakes context, where they could potentially land a job.

Interview Context and Individual Differences on User Behavior with an Agent. We next analyzed how interview context and individual differences influence users' behavior with an agent. To address our research questions, we were interested in two types of user behavior: (1) *willingness to confide* (IM scale as part of it) and (2) *willingness to listen*. To do so, we entered each measure as a dependent variable in a stepwise regression with interview context and individual differences as independent variables. Table 7 summarizes the results.

Willingness to confide. For Kaya, *Interview Context* and personality factors 1, 2, 3, 5, 6, and 7 (Table 5(a)) significantly impacted users' willingness to confide in Kaya. In particular, in a *high-stakes context*, *open-minded* and *cooperative* people were more likely to confide in Kaya. On the other hand, in such a context, more extroverted, excitement-seeking, anxious, and vulnerable people are *less* likely to confide in Kaya. It is worth noting that our results are consistent with the findings in social science studies, where *open-minded*, *agreeable* (cooperative) people are more

Table 7. Interview Context and Individual Differences Predicting User Behavior.

User Behavior w/ a REP	REP	Significant Factor	B(SE)	t	F (df)
Willingness to Confide	Kaya	Constant	1.78(.07)	26.19***	$F(7, 1087) = 145.22$ adj. $R^2 = .48$ $p < .0001$
		Interview Context	.76(.13)	5.78***	
		Factor 1 (Extroversion)	-.97(.06)	-16.57***	
		Factor 2 (Openness)	.26(.05)	4.99***	
		Factor 3 (Neuroticism)	-.30(.04)	-6.81***	
		Factor 5 (Trust)	-.32(.04)	-7.35***	
		Factor 6 (Cooperation)	.11(.04)	2.65**	
	Albert	n.s.			
IM Scale	Kaya	Constant	7.98(.18)	45.51***	$F(4, 1087) = 128.99$ adj. $R^2 = .32$ $p < .0001$
		Interview Context	3.73(.27)	13.94***	
		Gender	.54(.22)	2.44*	
		Factor 1 (Extroversion)	-.57(.13)	-4.37***	
		Factor 4 (Conscientious)	.35(.10)	3.39**	
	Albert	Constant	7.98(.18)	45.51***	$F(1, 151) = 4.79$ adj. $R^2 = .02$ $p < .03$
		Factor 2 (Neuroticism)	-.60(.28)	-2.19*	
Willingness to Listen	Kaya	Constant	3.18(.14)	22.20***	$F(2, 661) = 11.07$ adj. $R^2 = .03$ $p < .0001$
		Similar Personality	.23(.06)	4.06***	
	Albert	Factor 1 (Extroversion)	.23(.12)	2.00*	
	Albert	n.s.			

For the Kaya Dataset, the Independent Variables are Interview Context, Culture, Gender, Similar Personality, and Derived Personality Trait Factors Shown in Table 5(a). For the Albert Dataset, the Independent Variables are Gender, Similar Personality, and Derived Personality Trait Factors Shown in Table 5(b).

willing to trust strangers [29]. Considering Kaya a friendly stranger, users apply the same social rules to computers as found in many studies [68]. However, such trust did not show up in a low-stakes context. One reason might be that people were less engaged with an agent in that context. For Albert, no significant factors were found, which might be due to inadequate data (159 users).

IM Scale. As explained in Section 6.2, as part of the “willingness to confide” measure, the IM scale measures one’s tendency to inflate oneself by consciously lying during a self-evaluation and is often influenced by context [73] and individual differences [94]. On the surface, the IM scores were quite different under the two contexts of our deployments. In Deployments I and II for job interviews (high stakes), IM scores measured a $mean = 12.64$, $sd = 3.44$, while in Deployments III and IV in a university setting (lower stakes), they measured a $mean = 8.48$, $sd = 3.37$. Our step-wise regression showed that *Gender*, *Interview Context*, and personality factors 1 and 4 (Table 5(a)) significantly impact IM scores. In particular, in a high-stakes context, *conscientious males* tended to inflate their image much more than *extroverted females*. Our finding is consistent with previous results, which show that people who have high self-control (associated with *Conscientiousness*) score higher on the IM scale [94]. For Albert, only personality factor 2 (Table 5(b)) was significant, suggesting that emotionally *more* stable people inflated their image more.

Willingness to listen. *Similar Personality* and personality factor 1 (*Extroversion* and *Cheerfulness*) significantly affected users’ willingness to listen to Kaya. For Albert, however, no factors were significant. This implies that extroverted, cheerful users are more willing to listen to an agent

with a similar personality. Since Kaya was given a cheerful persona, the agreement between users' self-assessment (*Similar Personality* to Kaya) with their system-inferred traits (factor 1) seems to be validating the inferred personality traits. Moreover, our findings are consistent with other findings: users similar to an agent are more willing to cooperate with the agent [68, 86, 87].

9 SUMMARY OF FINDINGS AND DISCUSSION

Here, we summarize our findings and discuss the limitations in our current work, along with ongoing research, including design implications on building hyper-personalized intelligent agents based on automatically inferred user traits.

9.1 Findings and Implications

We summarize three key findings and state their implications on designing effective virtual interviewers.

Practical Values of AI Interviewers. Our field deployments demonstrated several practical values of AI interviewers. It makes a recruiting process more efficient, objective, and inclusive. In Deployment I, out of 800 applicants with half of them who simply ignored the interview invite and 59 who did not finish their interview, the hiring manager was able to quickly identify 12 quality candidates based on several inferred characteristics (e.g., *Responsible* and *Action-Oriented*) out of the 316 who completed their virtual interview. He also commented that a few of the 12 would have gone unnoticed by their resumes alone.

In Deployment II, 454 applicants signed up for an interview with Kaya, 307 started an interview, and 290 completed theirs. The hiring manager spent about 2 hours to select 60 candidates who ranked high on several dimensions (e.g., *Self-Discipline* and *Detail-Oriented*) and followed up with them in a technical interview to test their algorithmic and programming skills. Among the 60 candidates, 16 scored *Excellent* on a 4-level scale (*Weak, Ok, Good, and Excellent*). These results were impressive compared with a similar intern hiring process without the aid of an agent. In that process, the same hiring manager spent a couple of weeks to screen the resumes of over 400 applicants and selected 60 for the follow-up technical interview. Among the 60 students, only 6 of them were scored *Excellent*. In addition, the candidates selected based on their traits inferred from their virtual interview were much more diverse, including students from many different universities. Although we could not build a technical competence prediction model using the inferred personality traits for the small sample set ($N = 60$), just the amount of screening time saved and the diversity of the candidates discovered through the AI interviews make the agents quite valuable for hiring managers.

Furthermore, users appear to act authentically around an AI interviewer to reveal their true character even in a high-stakes situation. For example, one job applicant told Albert that his question was dumb, while another told Kaya to “*drink warm water*” after Kaya got stuck and said “*I just had a brain freeze*” when recovering from an error. Users also reported they “*felt more casual speaking with Kaya than an actual person, almost like writing in a journal,*” while others found the agent’s analysis very insightful, e.g., “*It was not something I had thought about before and appreciate the information.*” Perhaps one user’s comments to Kaya summarized well what an agent is built to achieve.

“I really enjoyed chatting with you. It was very nice getting to gain some insight on how someone else would view my personality without becoming judgmental.”

In addition to positive comments, users indicated aspects of agents to be improved. The top-two complaints were about (1) an agent’s response timing and (2) an agent’s lack of knowledge in its

responses. For example, one user complained that the agent's "*timing could be improved*" as the agent proceeded with the next question before he finished answering the previous one. After examining several chat transcripts, we observed that users often answer a question by sending multiple text messages. Besides detecting a user's keystrokes, currently, an agent cannot accurately predict whether and when a user has finished his answer to a given question. Another top complaint is that an agent's responses "are a bit too mechanical" or an agent "has limited knowledge of my responses." Overall, both critical comments indicate the current NLP limitations in our agents for handling diverse, free-text user responses in interviews similar to what was encountered by other systems [22].

Determining Personality of an AI Interviewer. Our findings indicate that an AI interviewer with Albert's personality and its matching behavior will make users willing to confide in and listen to the agent in a high-stakes situation, such as a job interview. However, in real-world deployments, business users are the ultimate decision-maker as to what kind of agent could be used for their target audience. We found out that business users often prefer a Kaya-like agent with a cheerful, friendly personality to engage with their young audience. From such uses, audience characteristics seem to be weighing more on the businesses' decisions.

Customizing an Agent Based on Users' Personality. Our extensive results show that users' personality traits inferred from their chat significantly impact their perceptions and behavior with an agent (Tables 6 and 7). It is thus desirable to customize an AI interviewer's behavior based on a user's inferred traits to make an interview more valuable and enjoyable. For example, our results show that in a high-stakes situation, achievement-striving users tend to trust an agent more and such trust also makes them eager to impress the agent. In such a case, an agent can remind such users to be honest, because they can achieve more with their true selves. Also, in many cases, if a user perceives an agent with a similar personality, this user tends to trust and favor the agent much more. For example, extroverted and emotional people are more willing to listen to an agent with a similar outgoing personality regardless the context. For such users, it may be the best to create an agent that matches their personality to make an interview more effective and enjoyable.

9.2 Limitations

Our current work also has several limitations due to the complex nature of real-world deployments as well as technology imperfections. Next, we enumerate the limitations and ongoing work to overcome such limitations.

Experiment Constraints. While our results were obtained from real-world deployments, this context posed a couple of limitations on the understanding of our results. First, our business users ultimately decided which AI interviewer to use in their context and thus we had little control on the type of data to be collected. For example, few business users chose to use Albert in their deployments. There is a lack of data for us to compare whether users would behave differently with Albert and Kaya in a less risky context, such as in our student surveys. Second, while previous work shows that gender and culture impact user interaction with virtual agents [26, 86], we did not find gender or culture to be a significant factor. This might be due to inadequate data or coarse metric such as how we differentiated users' cultural backgrounds. As we are deploying more agents to support diverse interview situations, it would be interesting to see whether our findings would hold across interview contexts.

While our field experiments have examined one's willingness to confide in and listen to agents, we have not studied in our context whether a user might overtrust an AI interviewer and what might be the consequences to the user. For example, researchers have already found that users

might over-trust robots in an emergency situation [80] or machine-inferred personality traits as ours [98]. To optimize human-machine symbiosis, we believe it is important to help users develop a balanced trust in machines.

Customizing an AI Interview. Currently, each interview question is represented as a conversation topic, and follow-up questions as sub-topics. Currently, our engine supports about 60 topics that handle conversations around 60 types of open-ended questions. On average, each topic includes about 15 rules. Typically, it would take a person who is familiar with our rule-based language (see rule examples in Figure 4(a)–4(b)) between 15–30 minutes to add and test a new topic, depending on the complexity of the topic (e.g., the diversity of user responses to be handled). Since our current topics have covered many major types of questions, it normally takes us about 2–4 hours to create and test a new interview containing 15–20 interview questions.

While it is straightforward to customize an interview by adding or removing interview questions as topics, it is impractical for a business user, such as a hiring manager, to do so manually. The task becomes even harder if an interview question is an open-ended question, which would then require a business user to consider possible user responses and how to handle them as sub-topics. Ideally, given a set of interview questions, we will automatically translate them into conversation topics. Toward this goal, we are examining patterns in interview questions before auto-translating them into topics.

Creating a Virtual Agent Beyond Interviewing. Theoretically, our topic-based conversation engine can be used to support virtual agents with varied capabilities by adding more rule-based topics. While our hybrid rule-based and algorithm approach has worked well for agenda-driven agents such as AI interviewers, it would not work well to support arbitrary user-driven conversations, such as an AI customer representative who must respond to diverse user requests. In these cases, it would be impractical to write a large number of rules to cover diverse user interaction patterns. Similar to building virtual agents like Apple Siri or Amazon Alexa, retrieval-based models trained on huge amounts of data might be more sensible [103] instead of using a rule-based approach as ours.

Inferring User Traits beyond Text. Currently, an agent infers a user's traits from the user's text given during the interview. However, we observed interesting user behavior beyond what their text captures during the whole process. For example, certain users asked for help (via email) when encountering technical difficulties (e.g., using an incompatible web browser) while others simply gave up (e.g., abandoning partially completed interviews). Moreover, during an interview, a user may respond to a question promptly but hesitate to reply to another question. In addition to a user's text, all such behavior could have been leveraged to build a better trait inference model. The challenge is how to capture such behavior, especially outside of an interview, and how to relate user behavior with specific traits.

9.3 Other Future Work

Here, we share a couple of research directions that are natural extensions to the current effort reported here.

Validating “Objectiveness” of Virtual Interviewers. When making an assessment or decision, research shows that humans often rely on intuitions not mentioning that human emotion is often a factor in such a process [6]. Studies also show that a pure data-driven approach beat human intuition at making an unbiased selection [52]. One of the reasons why we built AI interviewers is to render a more objective, unbiased assessment of interviewees than its human counterpart would do, especially in high-stakes contexts such as job interviews. While we have been very careful in

selecting our training data and algorithms, we have not quantitatively compared the objectivity of an AI interviewer versus that of a human interviewer. As part of our future work, we would like to do similar work as to that in Ref. [52], where we would measure objectivity against ground-truth (i.e., real-world outcomes such as a person's actual performance at workplace).

Design Implications Beyond Virtual Interviewers. Our findings indicate that inferred user traits significantly influence users' behavior with an agent. This implies that an agent could dynamically adapt its behavior to a user based on the user's traits inferred during a conversation. For example, when an *impulsive* user is unwilling to listen to a REP's guidance during a coaching session, the agent would try to calm him down first before continuing. On the other hand, an agent would encourage a *humble* person to open up and show her best talents. While much research recognizes the importance of incorporating user traits, such as personality, into an intelligent system [43, 90], few practical systems have been developed to obtain and use such traits in real time. Thus, automatically inferring user traits during a human-computer conversation enables the development of a new generation of intelligent agents who can truly understand their users on the fly and adapt to them in real time.

10 CONCLUSIONS

We have presented a new type of intelligent agent who can interview a user through a text-based conversation and automatically infer the user's traits, such as Big 5 personality traits and strengths. Our agents have been used in a number of field deployments and have interviewed thousands of users. In this article, through four real-world deployments of AI interviewers, we have examined how agent personalities and users' own personalities influence users' behavior—their willingness to confide in and listen to an agent during an interview in two very different contexts: a high-stakes job interview and a low-stakes educational setting. Our results from the four field deployments of five agents with 1,280 completed interviews have revealed several new findings. First, an AI interviewer can make a recruiting process more efficient, objective, and inclusive. Second, our results suggest that users seem to act quite authentically around an AI interviewer to reveal their true character. In particular, users are more willing to confide in and listen to a serious, assertive agent in a high-stakes situation. Third, users' own personality traits and their perceived personality similarity to an agent significantly influence their perception of and behavior with an agent. It is thus highly valuable to create hyper-personalized agents based on users' characteristics automatically inferred during a conversation.

ACKNOWLEDGMENTS

The authors would also like to thank Frances Thai and Connie Truong for their help in developing the dashboard.

REFERENCES

- [1] S. Adali and J. Golbeck. 2012. Predicting personality with social behavior. In *ASONAM'12*. 302–309.
- [2] N. Al Moubayed, Y. Vazquez-Alvarez, A. McKay, and A. Vinciarelli. 2014. Face-based automatic personality perception. In *Proceedings of ACM Multimedia 2014*. 1153–1156.
- [3] A. Aly and A. Tapus. 2013. A model for synthesizing a combined verbal and non-verbal behavior based on personality traits in human-robot interaction. In *ACM/IEEE Proceedings of HRI'13*. 325–332.
- [4] S. Argamon, S. Dhawle, M. Koppel, and J. Pennbaker. 2005. Lexical predictors of personality type. In *Proceedings of Interface and the Classification Society of North America*.
- [5] D. Arellano, J. Varona, F. Perales, N. Bee, K. Janowski, and E. Andre. 2011. Influence of head orientation in perception of personality traits in virtual agents. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems*. 1093–1094.

- [6] R. Baron. 1993. Interviewers' moods and evaluations of job applicants: The role of applicant qualifications. *J. Appl. Social Psychol.* 23, 4 (1993), 253–271.
- [7] A. Baylor and S. Ebbers. 2003. *The Pedagogical Agent Split-persona Effect: When Two Agents Are Better Than One*. EDMEDIA, Honolulu, Hawaii.
- [8] L. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe. 2011. Please, tell me about yourself: Automatic personality assessment using short self-presentations. In *Proceedings of the International Conference on Multimodal Interfaces*. 255–262.
- [9] T. Bickmore, A. Gruber, and R. Picard. 2005. Establishing the computer-patient working alliance in automated health behavior change interventions. *Patient Educ. Couns.* 59, 1 (2005), 21–30.
- [10] T. Bickmore and J. Cassell. 1999. Small talk and conversational storytelling in embodied interface agents. In *Proceedings of the AAAI Fall Symposium on Narrative Intelligence*.
- [11] F. Bouchet and J. Sansonnet. 2012. Intelligent agents with personality: From adjectives to behavioral schemes. In *Cognitively Informed Interfaces: System Design and Development*, E. M. Alkhalifa and K. Gaid (Eds.). IGI Global, 177–200.
- [12] P. Brown and S. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press.
- [13] A. Cafaro et al. 2012. First impressions: Users' judgments of virtual agents' personality and interpersonal attitude in first encounters. In *Proceedings of the International Conference on Intelligent Virtual Agents*. Springer, Berlin.
- [14] R. Cialdini. 2006. *Influence: The Psychology of Persuasion*. Harper Business.
- [15] Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Routledge.
- [16] L. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 3 (1951), 297–334.
- [17] P. T. Costa Jr. and R. R. McCrae. 1992. *Revised NEO Personality Inventory (Neo-PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional manual*. Psychological Assessment Resources, Odessa, FL.
- [18] N. Draper and H. Smith. 1981. *Applied Regression Analysis*, 2nd ed. John Wiley & Sons, Inc., New York.
- [19] R. J. de Ayala. 2009. *The Theory and Practice of Item Response Theory*. Guilford Press.
- [20] C. DeYoung, L. Quilty, and J. Peterson. 2007. Between facets and domains: 10 aspects of the Big Five. *J. Personality Social Psychol.* 93, 5 (2007), 880–896.
- [21] A. Dempster, N. Laird, and D. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.* 39, 1 (1977), 1–38.
- [22] D. DeVault, R. Benn Artstein, G. Dey, E. R. Fast, A. Gainer, K. Georgila, D. Gratch, A. Hartholt, A. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, S. Stratou, A. Suri, D. Traum, W. Wood, Y. Xu, A. Rizzo, and L. P. Morency. 2014. SimSensei kiosk: A virtual human interviewer for healthcare decision support. In *AAMAS'14*.
- [23] J. M. Digman. 1990. Personality structure: Emergence of the five-factor model. *Annu. Rev. Psychol.* 41, 1 (1990), 417–440.
- [24] D. Christopher Dryer. 1999. Getting personal with computers: How to design personalities for agents. *J. Appl. Artif. Intell.* 13, 3 (1999), 273–295.
- [25] S. Escalera, V. Athitsos, and I. Guyon. 2016. Challenges in multimodal gesture recognition. *J. Mach. Learn. Res.* 17 (2016), 1–54.
- [26] B. Endrass, E. André, M. Rehm, and Y. Nakano. 2013. Investigating culture-related aspects of behavior for virtual characters. *J. Auton Agents Multi-Agent Syst.* 27, 2 (2013), 277–304.
- [27] J. Fan, D. Gao, A. Carroll, S. Lopez, J. Siva, T. Tian, and H. Meng. 2012. Testing the efficacy of a new procedure for reducing faking on personality tests within selection contexts. *J. Appl. Psychol.* 97 (2012), 866–880.
- [28] B. J. Fogg and C. Nass. 1997. Silicon sycophants: The effects of computers that flatter. *Int. J. Hum. Comput. Stud.* 46, 5 (1997), 551–561.
- [29] M. Freitag and P. Bauer. 2016. Personality traits and the propensity to trust friends and strangers. *Social Sci. J.* 53, 4 (2016), 467–476.
- [30] P. Gebhard, T. Baur, I. Damian, G. Mehlmann, J. Wagner, and E. André. 2014. Exploring interaction strategies for virtual characters to induce stress in simulated job interviews. In *AAMAS'14*. 661–668.
- [31] A. Gill, S. Nowson, and J. Oberlander. 2009. What are they blogging about? Personality, topic and motivation in blogs. In *ICWSM'09*. 18–25.
- [32] M. Gludice, T. Booth, and P. Irwing. 2012. The distance between Mars and Venus: Measuring global sex differences in personality. *Pub. Lib. Sci.* 7, 1 (2012).
- [33] L. Goldberg. 1981. Language and individual differences: The search for universals in personality lexicons. In *Review of Personality and Social Psychology*, vol. 2. 141–165.
- [34] B. Grosz and C. Sidner. 1986. Attention, intention, and the structure of discourse. *Comput. Ling.* 12, 3 (1986), 175–204.
- [35] L. Gou, M. X. Zhou, and H. Yang. 2014. *KnowMe and ShareMe: Understanding Automatically Discovered Personality Traits from Social Media and User Sharing Preferences (CHI'14)*. 955–964.
- [36] L. Havens. 1986. *Making Contact: Uses of Language in Psychotherapy*. Harvard University Press.

- [37] L. Helgoe. 2013. *Introvert Power: Why Your Inner Life is Your Hidden Strength*, 2nd ed. Sourcebooks.
- [38] W. Hofstee, B. Raad, and L. Goldberg. 1992. Integration of the Big Five and circumplex approaches to trait structure. *J. of Personality Social Psychol.* 63, 1 (1992), 146–163.
- [39] J. Holland. 1985. *Making Vocational Choices: A Theory of Vocational Personalities and Work Environments*. Prentice-Hall, Engelwood Cliffs, NJ.
- [40] J. Hopcroft. 1971. *An Nlogn Algorithm for Minimizing States in a Finite Automaton, Theory of Machines and Computations*. Academic Press, 189–196.
- [41] J. Hopcroft and J. Ullman. 2006. *Introduction to Automata Theory, Languages, and Computation*, 3rd ed. Pearson.
- [42] D. Hoover. 2003. Another perspective on vocabulary richness. *Comput. Humanities* 37 (2003), 151–178.
- [43] R. Hu, P. Pu. 2011. Enhancing collaborative filtering systems with personality information. In *RecSys’11*.
- [44] A. V. Ivanov, G. Riccardi, A. J. Sporka, and J. Franc. 2011. Recognition of personality traits from human spoken conversations. In *Proceedings of InterSpeech*. 1549–1552.
- [45] A. Jackle, P. Lynn, J. Sinibaldi, and S. Tipping. 2013. The effect of interviewer experience, attitudes, personality, and skills on respondent cooperation with face-to-face surveys. *Surv. Res. Methods* 7, 1 (2013), 1–15.
- [46] H. Jones, N. Sabouret, I. Damian, T. Baur, E. André, K. Porayska-Pomsta, and P. Rizzo. 2014. Interpreting social cues to generate credible affective reactions of virtual job interviewers. CoRR abs/1402.5039.
- [47] M. Kern, J. Eichstaedt, A. Schwartz, L. Dziurzynski, L. Ungar, D. Stillwell, M. Kosinski, S. Ramones, and M. Seligman. 2013. The online social self: An open vocabulary approach to personality. *J. Assess.* 21, 2 (2013), 158–169.
- [48] T. Koda and P. Maes. 1996. Agents with faces: The effects of personification of agents. In *Proc. HCI’96*. 98–103.
- [49] M. Kosinski, D. Stillwell, and T. Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proc. Natl. Acad. Sci. USA* 110, 15 (2013), 5802–5805.
- [50] M. Kubat and J. Milicka. 2013. Vocabulary richness measure in genres. *J. Quant. Ling.* 20, 4 (2013), 339–349.
- [51] V. Kulkarni, M. Kern, D. Stillwell, M. Kosinski, S. Matz, L. Ungar, S. Skiena, and A. Schwartz. 2017. Latent human traits in the language of social media: An open-vocabulary approach. In *CORR*, <http://arxiv.org/abs/1705.0803>.
- [52] N. Kuncel, D. Klieger, B. Connelly, and D. Ones. 2013. Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *J. Appl. Psychol.* 98, 6 (2013), 1060–1072.
- [53] J. E. Laird. 2012. *The Soar Cognitive Architecture*. MIT Press.
- [54] K. Lee, W. Peng, S. Jin, and C. Yan. 2006. Can robots manifest personality? An empirical test of personality recognition, social responses, and social presence in human-robot interaction. *J. Comm.* 56, 4 (2006), 754–772.
- [55] K. Lee, J. Mahmud, J. Chen, and M. X. Zhou, and J. Nichols. 2014. Who will retweet this? In *Automatically Identifying and Engaging Strangers on Twitter to Spread Information (IUT’14)*. 247–256.
- [56] B. Lepri, R. Subramanian, K. Kalimeri, J. Staiano, F. Pianes, and N. Sebe. 2012. Connecting meeting behavior with extraversion: A systematic study. *IEEE Trans. Affective Comput.* 3, 4 (2012), 443–455.
- [57] S. Li and M. Roloff. 2006. Strategic emotion in negotiation: Cognition, emotion, and culture. In *Communication to Presence: Cognition, Emotions and Culture towards the Ultimate Communicative Experience*. 166–185.
- [58] L. Liu, D. Preotiu-Pietro, Z. Riahi Samani, M. Moghaddam, and L. Ungar. 2016. Analyzing personality through social media profile picture choice. In *ICWSM’16*.
- [59] G. Lucas, J. Gratch, A. King, and L. Morency. 2014. It’s only a computer: Virtual humans increase willingness to disclose. *Comp. Hum. Behav.* 37 (2014), 94–100.
- [60] E. Luger and A. Sellen. 2016. “Like having a really bad PA”: The gulf between user expectation and experience of conversational agents. In *ACM SIGCHI’16*.
- [61] L. Luo, F. Wang, M. X. Zhou, Y. Pan, and H. Chen. 2014. Who have got answers? Growing the pool of answerers in a smart enterprise social QA system. In *IUT’14*. 7–16.
- [62] J. Mahmud, M. X. Zhou, N. Megiddo, J. Nichols, and C. Drews. 2013. Recommending targeted strangers from whom to solicit information on social media. In *IUT’13*. 37–48.
- [63] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *JAIR* 30 (2007), 457–500.
- [64] R. C. Mayer, J. H. Davis, and F. D. Schoorman. 1995. An integrative model of organizational trust. *Acad. Manage. Rev.* 20, 3 (1995), 709–734.
- [65] M. McRorie, I. Sneddon, G. McKeown, E. Bevacqua, E. de Sevin, and C. Pelachaud. 2011. Evaluation of four designed virtual agent personalities. *IEEE Trans. Affective Comput.* 3, 3 (2011), 311–322.
- [66] G. Mehlmann, K. Janowski, and E. André. 2016. Modeling grounding for interactive social companions. *KI* 30, 1 (2016), 45–52.
- [67] M. Mohri. 1997. Finite-state transducers in language and speech processing. *Comput. Ling.* 23, 2 (1997), 269–311.
- [68] C. Nass and C. Yen. 2012. *The Man Who Lied to His Laptop*. Current.
- [69] C. Nass, J. Steuer, and E. Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 72–78.

- [70] K. Nowak, J. Fox, and Y. Ranjit. 2015. Inferences about avatars: Sexism, appropriateness, anthropomorphism, and the objectification of female virtual representations. *J. Comput. Mediated Commun.* 20, 5 (2015), 554–569.
- [71] J. Nunamaker et al. 2011. Embodied conversational agent-based kiosk for automated interviewing. *J. Manage. Inf. Syst.* 28, 1 (2011), 17–48.
- [72] B. Okun. 2007. *Effective Helping: Interviewing and Counseling Techniques*, 7th ed. Cengage Learning.
- [73] D. L. Paulhus. 1991. Measurement and control of response bias. In *Measures of Personality and Social Psychological Attitudes*, J. P. Robinson and P. R. Shaver (Eds.). Vol. 1. Academic Press, 17–59.
- [74] J. Pennebaker and L. King. 1999. Linguistic styles: Language use as an individual difference. *J. Personality Social Psychol.* 77, 6 (1999), 1296–1312.
- [75] S. Piana, A. Staglianò, F. Odone, and A. Camurri. 2016. Adaptive body gesture representation for automatic emotion recognition. *ACM TITS* 6, 1 (2016), 1–31.
- [76] D. Preotiuc-Pietro, J. Carpenter, S. Giorgi, L. Ungar. 2016. Studying the dark triad of personality through Twitter behavior. In *CIKM'16*. 761–770.
- [77] A. Powers, S. Kiesler, and J. Goetz. 2003. Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *Proceedings of the 12th IEEE International Workshop on Robot and Human Interactive Communication*, vol. IXX, 55–60.
- [78] B. Reeves and C. Nass. 2002. *The Media Equation*. CSLI Publications.
- [79] T. Rath. 2007. *Strengths Finder 2.0*. Gallup Press.
- [80] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *HRI'16*. 101–108.
- [81] J. Sansonnet and F. Bouchet. 2013. Managing personality influences in dialogical agents. In *Proceedings of ICAART'13*, Vol. 1. 89–98.
- [82] M. Schmitz, A. Kruger, and S. Schmidt. 2007. Modeling personality in voice of talking products through prosodic parameters. In *Proceedings of the International Conference on the Internet of Things*. 313–316.
- [83] J. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.
- [84] I. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI'16*. 3776–3783.
- [85] L. Shao, S. Gouws, D. Britz, A. Goldie, B. Strophe, and R. Kurzweil. 2017. Generating long and diverse responses with neural conversational models. In *ICLR'17*.
- [86] L. Sproull, M. Subramani, S. Kiesler, J. Walker, and K. Waters. 1996. When the interface is a face. *J. Hum. Comput. Interact.* 11, 2 (1996), 97–124.
- [87] A. Tapus, C. Tapus, and M. Mataric. 2008. User-robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy. *Intell. Serv. Rob.* 1, 2 (2008), 169–196.
- [88] Y. Tausczik and J. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Social Psychol.* 29, 1 (2010), 24–54.
- [89] P. Thiel and B. Masters. 2014. *Zero to One: Notes on Startups, or How to Build the Future*. Currency.
- [90] N. Tintarev, M. Dennis, and J. Masthoff. 2013. Adapting recommendation diversity to openness to experience: A study of human behavior. In *UMAP'13*. 190–202.
- [91] D. Traum. 2017. Computational approaches to dialogue. In *The Routledge Handbook of Language and Dialogue*, E. Weigand (Ed.), New York.
- [92] J. Trouvain, S. Schmidt, M. Schroder, M. Schmitz, and W. J. Barry. 2006. Modeling personality features by changing prosody in synthetic speech. In *Proceedings of Speech Prosody*.
- [93] M. Turk. 2014. Multimodal interaction: A review. *Pattern Recognit. Lett.* 36 (2014), 189–195.
- [94] L. Uziel. 2013. Impression management (“lie”) scales are associated with interpersonally oriented self-control, not other deception. *J. Personality* 82, 3 (2013), 200–212.
- [95] A. Vinciarelli and G. Mohammadi. 2014. A survey of personality computing. *IEEE Trans. Affective Comput.* 5, 3 (2014), 273–291.
- [96] O. Vinyals and Q. Le. 2015. A neural conversational model. In *Proceedings of ICML, Deep Learning Workshop*.
- [97] R. Wallace, H. Tomabechi, and D. Aimless. 2007. Chatterbots Go Native: Considerations for an eco-system fostering the development of artificial life forms in a human world. Retrieved from <http://www.pandorabots.com/pandora/pics/chatterbotsgonative.doc>.
- [98] J. Warshaw, T. Matthews, S. Whittaker, C. Kau, M. Bengualid, and B. Smith. 2015. Can an algorithm know the real you? Understanding people’s reactions to hyper-personal analytics systems. In *CHI'15*. 797–806.
- [99] Y. Weisberg, C. DeYoung, and J. Hirsh. 2011. Gender differences in personality across the ten aspects of Big Five. *J. Front Psychol.* 2, Article 178 (2011), 11 pages. <http://dx.doi.org/10.3389/fpsyg.2011.00178>
- [100] J. Weizenbaum. 1966. ELIZA—A computer program for the study of natural language communication between man and machine. *Commun. ACM* 9 (1966), 36–45.

- [101] A. Wexelblat. 1998. Don't make that face: A report on anthropomorphizing and interface. In *AAAI'98*. 173–179.
- [102] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke. 2017. The Microsoft 2017 Conversational Speech Recognition System, Tech Report, Retrieved from <https://www.microsoft.com/en-us/research/publication/microsoft-2017-conversational-speech-recognition-system/>.
- [103] A. Xu, Z. Liu, Y. Guo, V. Sinha, R. Akkiraju. 2017. A new chatbot for customer service on social media. In *Proc. ACM CHI'17*. 3506–3510.
- [104] T. Yarkoni. 2010. Personality in 100,000 words: A large-scale analysis of personality and word usage among bloggers. *J. Res. Personality* 44, 3 (2010), 363–373.
- [105] N. Yee, J. Bailenson, and K. Rickertsen. 2007. A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1–10.
- [106] R. Yu, Y. Liu, and M. Yang. 2011. Does interviewer personality matter for survey outcomes? In *Proceedings of the 64th Annual Conference on World Association for Public Opinion Research*. 89–118.
- [107] R. B. Zajonc. 1980. Feeling and thinking: Preferences need no inferences. *Am. Psychologist* 35, 2 (1980), 151–175.
- [108] M. Ziegler, C. MacCann, and R. Roberts, (Ed). 2010. *New Perspectives on Faking in Personality Assessment*. Oxford University Press.

Received June 2017; revised September 2018; accepted September 2018