Evaluation of Four Designed Virtual Agent Personalities

M. McRorie[1], I. Sneddon[1], G. McKeown[1], E. Bevacqua[2], E. de Sevin[3] and C. Pelachaud[2]

[1]School of Psychology, Queen's University Belfast, Belfast, United Kingdom
[2]CNRS - LTCI UMR 5141, Institut TELECOM - TELECOM ParisTech, Paris, France
[3]LIP6 UMPC, Paris

Abstract—Convincing conversational agents require a coherent set of behavioural responses that can be interpreted by a human observer as indicative of a personality. This paper discusses the continued development and subsequent evaluation of virtual agents based on sound psychological principles. We use Eysenck's theoretical basis to explain aspects of the characterization of our agents, and we describe an architecture where personality affects the agent's global behaviour quality as well as their backchannel productions. Drawing on psychological research, we evaluate perception of our agents' personalities and credibility by human viewers (N=187). Our results suggest that we succeeded in validating theoretically grounded indicators of personality in our virtual agents,and that it is feasible to place our characters on Eysenck's scales. A key finding is that the presence of behavioural characteristics reinforces the prescribed personality profiles that are already emerging from the still images. Our long-term goal is to enhance agents' ability to sustain realistic interaction with human users, and we discuss how this preliminary work may be further developed to include more systematic variation of Eysenck's personality scales.

Index Terms—Personality traits, Eysenck, emotional traits, virtual agents

1 Introduction

We are all familiar with people in our daily lives demonstrating relatively stable, and often predictable, sets of behavioural characteristics. From such perceptions we automatically make judgements about the personalities of those we interact with. We need to be able to make the same kind of judgements about virtual agents. The credibility of such agents is dependent on them being perceived as coherent entities. To date, many of the behavioural characteristics that we use to make such automatic judgements (e.g. facial expressions, head and eye movements etc.) have been added to virtual agents in a way that is based, at best, on intuition. However, as we move toward a situation in which virtual agents are required to sustain extended durations of single interactions with users, the believability of such intuitively based ad hoc characters is likely to break down. We propose that it is essential for virtual agents to display physical appearance and behavioural characteristics that are sufficiently coherent to allow users to make the same kinds of inferences about personality that they continually make about the people around them in their daily lives. We anticipate that the coherence across behavioural characteristics generated in this way will add depth to people's perception of the characters – thus sustaining the 'believability' of the character over longer periods of time.

Our work is part of the European project SEMAINE[1], which aims to provide a multimodal system that allows interaction with conversational agents, or Sensitive Artificial Listeners (SAL). These virtual agents are designed to sustain realistic interaction with human users, despite having limited verbal skills. In this case we are not building these characters

[1]Schröder, M., "SEMAINE Project," Apr. 2011; http://www.semaine-project.eu/

from scratch. The agents have gone through several developmental stages as the role of the human operator has decreased and the autonomy of the agent has increased. In the early stages of this evolution, the appearance, behaviour and dialogue content were largely based on the intuition of the designers [1]. However if our agents are to demonstrate psychologically plausible behaviour, a theoretically sound approach to character development is required. Within the context of human-machine interaction, one of the most important features of a believable agent is a distinct personality [2]. Furthermore, research suggests that human users interpret the behaviour of virtual agents using the same social rules which are used to understand people [3]. Agentsare considered "believable" when they are perceived to portray the qualities and behaviourstypical of different personality types. For the purposes of this study, we thus consider believability and personality to be largely inter-connected, with genuine believability of an avatar depending on the perception of a personality. Trait models of personality assume that traits influence behaviour, and that they are stable, fundamental properties of an individual. In addressing the issue of believability in virtual agents, Ortony [4, p.202] similarly refers to the need for consistent and coherent (i.e. believable) characters, and he argues that production of truly believable agents will require using personality as a 'generative engine' which contributes to the coherence, consistency and predictability of their behavioural responses.

Currently we are attempting to shape the continuing development of the agents in ways that are more consistent with existing psychological knowledge about the links between, on the one hand, physical appearance and behaviour and, on the other, judgements about personality. Based on the same sound psychological framework, four distinct characters (SAL agents) have been created - each employing individual dialogue strategies, and displaying different reactions. Full explanation of how we selected Eysenck'stheory of personality and comprehensive information on the continuing development of our agents with different behaviour propensities are presented in [5]. This current paper provides a brief overview of how our choice of theory allowed us to remain computationally tractable yet still retain a realistic level of complexity to influence personality in virtual agents. A summary of the SAL architecture is also provided, where we outline how personality influences not only the behavioural characteristics of the virtual agents, but also their communicative styles. We then focus on our recent work undertaken to evaluate perception of agents' personality and credibility by human viewers.

1.1 Selecting Eysenck's theory

Psychological research on personality attribution tends to use one of the two main theories (five-factor [6] or three-factor model [7]). The five-factor model is a modern lexical (language based) approach. It is one of the most widely used trait theories and posits five main, relatively independent personality dimensions: extraversion, neuroticism, openness to experience, agreeableness and conscientiousness. In comparison, Eysenck developed a model based on traits which he believed were heritable and had a probable biological foundation. The three main traits which met these criteria were extra-

version-introversion, neuroticism-emotional stability, and psychoticism. Eysenck's dimensions of extraversion and neuroticism are virtually identical to the similarly named traits of the five-factor model and psychoticism corresponds to low agreeableness and low conscientiousness combined [8].

There is a continued debate in the literature [9] as to which of the two main models is more theoretically appropriate for understanding human characteristics. In addition it has been argued that virtual agents need easily controlled parameters [10], and for the purposes of this study we deemed it appropriate to start with the simpler of the two trait models to explore whether a convincing character could be developed based on three rather than five dimensions. We also wanted to avoid the strongly lexical foundation on which the five-factor model is based. The key point however is that Eysenck attempted to provide not merely a description of personality, but an explanation of cause. The merits of adopting this type of approach are that its biological underpinnings can provide a starting point for generating specific response patterns of behaviour – and thus believability - in virtual agents.

2 BUILDING PERSONALITY

Our objective was to provide a sound theoretical basis to generate behavioural characteristics which should allow an observer to infer adefinedpersonality. Personality predicts specific behaviours, and individual personality types are deduced from personality questionnaires, i.e. the self-reported answers to questions about behaviours. In developing credible artificial agents we needed to move in the opposite direction and generate consistent sets of behavioural attributes from personality theory.

2.1 Modeling agents with distinctive behavioural characteristics

Agents' behavioural tendencies were modelled using the approach developed by [11] where an agent is defined by a *baseline*, which captures the agent's global behaviour tendency in terms of the preference the agent has in using a modality (head, gaze, face, gesture, and torso) and on the expressive quality of each of these modalities. This baseline is defined as a set of numeric parameters. *Modality preference* refers to the agent's degree of preference in using each available modality to communicate, whereas the *behaviour expressivity* is represented by the set of 6 parameters (frequency, speed, spatial volume, energy, fluidity, and repetitivity) for each modality which influences the quality of the agent's movements as proposed by [12]. Thus, a baseline contains 35 parameters (1 degree of preference and 6 expressivity parameters for each one of the 5 modalities considered in our system). For example, through the baseline we can specify an agent that conveys information mainly with its face, gaze, gesture, or head movement and that has the tendency to move slowly, or in a faster manner on these modalities. Table 1 shows the baseline for each SAL agent for the two modalities of 'face' and 'gesture'.

2.2 Defining behaviour of SAL characters

Our next challenge was to consider how to translate stable traits into personality-dependent behavioural characteristics. Within the psychological literature, the major categories typically used to classify nonverbal behaviour are facial expressions, eye and visual behaviour (e.g. gaze), and paralanguage. By identifying and setting these types of parameters, it is possible to equip the agents with specific documented behaviour characteristics [11], [12], [13], [14], which relate to personality traits.

Although the literature describing behaviours associated with particular human personalities is not couched in the same terms used to describe the agents, we can begin by using the human research to help us specify the development parameters for our characters. The application consists of a system of Sensitive Artificial Listeners (SAL), designed to sustain a conversational interaction with a human user via generation of nonverbal behaviour in real time. Four psychologically different personality types have been created, each trying to draw the human user into their own emotional state: Poppy is outgoing (extraverted) and optimistic; Spike is angry and argumentative; Obadiah is gloomy and depressed; and Prudence is pragmatic and practical. Fig 1 portrays the four SAL facial models.

As explained in the following, we have associated sets of physical appearance and behavioural characteristics to each of the SAL personality types. Whilst the agents have not been provided with every attribute associated with their intended personality trait, each agent is imbued with some of these characteristics.

Physical appearance. Poppy has been given an attractive appearance and a friendly facial expression, because we want people to attribute positive personality characteristics. The facial expressions of extraverts tend to be friendly [15], with positive personality attributions likely to be projected on to those possessing attractive faces [16]. Spike's facial features have been set in a permanently angry configuration because he exhibits hostile behaviour so frequently. This agent's dispositional qualities of being angry and argumentative relate to psychoticism, which involves elements of aggression, coldness and impulsivity. Facial expressions of anger are demonstrated with frowning eyebrows and staring eyes [17], with increased facial threat typified via prolonged direct eye gaze and wide eyes [18]. Prudence has been given a symmetrical facial appearance, her hair is pulled back in a business-like manner, and she wears glasses. Designed to appear practical and pragmatic, Prudence's defining characteristics suggest conscientiousness, thus indicating low impulsivity and low psychoticism. Faces high in symmetry have received significantly higher ratings for competence, intelligence and agreeableness [19]; and individuals wearing glasses tend to be rated as more intelligent [20]. Obadiah's defining features are gloominess and depression, which are characteristic of neuroticism, the tendency to experience negative emotional states. Negative facial expression is directly related to neuroticism [21].

Behavioural characteristics. Eysenck [7] proposed that individual differences in nervous system structure/functioning

could account for the emergence of personality traits. The biological basis of extraversion suggests extraverts are less cortically aroused than introverts. Drawing on Hebb's [22] notion of optimal level of arousal, this implies that extraverts should be more comfortable under arousing conditions. Poppy is thus characterized as having high levels of general activation. Extraverts tend to demonstrate more body movements, and display greater levels of facial activity [23]. Studies have also shown that extraversion is associated with greater levels of gesturing, more frequent head nods, and general speed of movement [15]. When communicating, extraverts are more likely to maintain direct facial posture and eye contact [24]. Extraverts tend to demonstrate fewer pauses, shorter silences, and fewer hesitations [25]. Spike is designed to display impulsive, aggressive qualities. Eysenck proposed that psychoticism - like extraversion - reflectslow cortical arousal, but is linked to levels of male hormones (e.g. testosterone) that influence impulsivity. Individuals high in psychoticism tend to be verbally aggressive, argumentative and inappropriately assertive in communication [25]. When communicating, high scorers on disagreeableness display less visual attention, but more visual dominance. Disagreeable individuals do less back-channelling, indicating they listen less to conversational partners [26]. On the other hand Prudence has been given behavioural characteristics that the literature suggests are associated with low levels of psychoticism. For example, individuals who are thoughtful and reflective may show a predominance of upward looks [27], and high eye contact has been linked to competence, confidence, and self-esteem. Conscientious characters such as Prudence tend to avoid negations, negative emotion words and words reflecting discrepancies (e.g. should and would). Obadiah is created with behaviour that can be associated with some aspects of neuroticism. This agent's speech tends to have low variation and a rather flat tone that reflects an emotional state which is low in activation. The literature suggests neuroticism predicts a negative emotional tone, and when communicating, high neuroticism scorers tend to have low, constant voice intensity [28]. Gaze avoidance and less eye contact are further cues [29].

3 SAL ARCHITECTURE

Our system uses the SEMAINE API, a distributed multi-platform component integration framework for real-time interactive systems [30]. User's acoustic and visual cues are extracted by analyser modules and then interpreted to derive information about the user (e.g. her emotional state and behavioral activity) and the dialogue's state (such as change of speaking turn). The next step is to decide whether the agent should act either as a speaker or a listener. When the agent is the speaker, the Dialogue Manager module determines which sentence the agent can utter and with which communicative functions. Sentences are selected from a set of predefined utterances specific to each agent's personality trait [35]. The communicative functions are instantiated in a list of multimodal behavioural signals. All possible sets of behaviours for a given communicative function are defined in the agent's lexicon. On the other hand, when the agent is the listener, the

Listener Intent Planner module decides when and how it should provide a backchannel signal. A two-steps algorithm is implemented. First potential backchannels are detected [5]. This step is done by analysing the user's acoustic and visual behavior [31; 32]. As a second step, these potential backchannels are filtered by the Backchannel Selection module. It computes the displayed backchannels according to agent's personality traits (emotion stability and extraversion) [61].

Each SAL agent is associated with its lexicon and its baseline. The lexicon for an agent contains the set of behaviours for any given communicative function. Each lexicon has been determined partly through perceptive tests [33; 34] and partly by analysing videos of human interactions from the SEMAINE database [35]. Each baseline is also set through videos analysis. For each agent, our system uses its own lexicon and expressivity parameters to compute the displayed behaviors. While it is not a model of personality per se, our approach does allow us to model agents with behavior patterns linked to its personality traits. Every time a character is called to interact with the user, both its lexicon and its baseline are automatically switched. In this way the generated animation varies according to the active agent.

Afterwards, the agent behaviour is realised according to the agent's baseline characteristics that corresponds to the modality preference (gesture, face, head) and expressivity values on each modality [11]. Finally, the agent's animation is rendered and displayed on a PC screen in real-time.

4 Evaluation of Agent Personality by Human Viewers

Four distinctive agents have been created, each according to a specificpersonality. Moreover, we have developed a system that allows us to modify the values of the agents' parameters in real time. This effort has been guidedby psychological theory, with the aim to enhance the believability of the agents - albeit for psychologists, the issue of believability has not been important up until this point. Drawing again on personality literature, we adopted a method of testing our system which is frequently used when psychologists focus on individual differences between people. We evaluated agents' personality and credibility as perceived by human viewers.

4.1 Personality judgements

Our behavioural characteristics are an important indicator of personality, and tend to be consistent across time and situations [36]. It is thus possible to make broad judgements about the behavior of others and then use these judgements to predict future behavior. Our initial impressions strongly influence subsequent expectations about others' behaviour and show surprising levels of inter-observer consensus and accuracy [37]. Consensus and accuracy in judging personality are important topics of research in psychology [38]. The most common way to assess accuracy is self-other agreement. Some studies suggest that self-other agreement increases with acquaintance [39] and that consensus of personality ratings of

close acquaintances is higher than ratings based on 'zero acquaintance' (observations of strangers). Others report no significant differences between the two [40]. Whilst the evidence is mixed, it is nonetheless clear that exposure to short observations of a target's behaviour can yield significant self-stranger agreement [37]. Within the current context, this suggests that people should be capable of inferring personality based on zero acquaintance of virtual agents. We have used this approach to explore whether viewers' (i.e. 'strangers') perception of agent's character is consistent with the agent's 'actual' personality as we have intended.

Methods of measuring personality. There is a broad literature which confirms that personality ratings based on reliable and valid personality questionnaires are repeatedly consistent between raters, and can demonstrate strong positive correlations between raters and self-reported personality [41]. Scores from adjective based scales such as the First Impression Interaction Procedure also provide consensus with self-ratings [42].

The stimuli used in such work can consist of short observations of behaviour, video clips or static facial images. The nature of the research tends to dictate the choice of stimuli employed. Close acquaintance research clearly depends on actual social contact. In zero acquaintance studies, short video clips are commonly used as stimuli, and a number of authors have demonstrated that brief expressive behaviour provides useful information [see meta-analysis, 43]. Other researchers opt for static photographs, and argue that the use of video footage which displays clothing, context etc may allow extraneous information to influence personality judgement [44]. The argument for using static images is that a viewer's attention has to be focused explicitly on information provided by the target person's innate facial features, and there is evidence that people frequently make personality judgements based on these cues alone [45].

Personality ratings based purely on facial appearance are not the focus of the current study. We provided our agents with a range of behavioural characteristics which we hoped would be perceived by raters as personality-dependent actions. We have constructed virtual characters varying in visual appearance, voice quality, vocal content and tone, and nonverbal behaviour. Vocal content was designed to reflect personality, and is part of agents' behavioural elements repertoire. Selection of voice was by consensusof an international team of emotion/affective science researchers), and based on perceived congruence between actors and scripted content for each agent. One focus of the current study was to evaluate the importance of these behavioural and non-behavioural elements influencing the personality judgements of perceivers.

Evaluation study. Our ultimate goal is to evaluate real-time interaction by human users. However there is a trade-off between the enhanced ecological validity of real-time interaction, and the heightened control provided by showing the same sequence of communication and behavior to all participants. The realities of providing *comparable* real-time interactions

with agents require consistent conversational interactions, and in this first stage of our research we opted for control. A further objective was to ensure raters' attention was focused solely on agent characteristics. When people are involved in real-time interactions, they may be distracted by the external and internal demands of the interaction itself, and research suggests that people involved in face-to face interactions may not be as accurate as those who observe video extracts of interactive behaviour [46]. Based on the zero-acquaintance paradigm we thus used a 'first impressions' approach. We believe this is an important first step in evaluating the personalities (and thus believability) of our agents.

In order to assess the influence of each Big Three dimension on perception of our characters, personality was rated both on the basis of facial appearance alone (static coloured pictures), and on behavioural characteristics (short video clips of interactions between agents and human users). We also conducted adjectival analyses for each character, gauging the prevalence of adjectives typically associated with each of the three personality types. Supplementary ratings of believability, familiarity and consistency facilitated further insight in respect of coherence and credibility. It was predicted that Poppy would be rated highest in extraversion, with Spike perceived to be highest in psychoticism. We anticipated that Obadiah would obtain the highest neuroticism score. Prudence was created to be practical and pragmatic, with her defining characteristics expected to be conscientiousness. We thus expected this agent to obtain the lowest psychoticism score.

Given that facial appearance alone can be an accurate predictor of personality, we included comparisons of mean ratings for static and moving images. This allowed us 1) to assess the value of our work based on facial features alone, and 2) more fundamentally, to tease out the comparative richness (if any) of the information provided by the behavioural sequences. If we have been successful in creating agent personalities, then more information should be available for the viewer from the dynamic moving images, and mean ratings for each agent's defined trait should be higher.

In selecting dynamic stimuli we addressed issues regarding exposure length and location extraction. The term 'thin slices' is commonly used in human personality research, and refers to short extracts of behaviour from which viewers can make judgements about personality traits and affective states [43]. It is an approach which is ideal for considering first impressions, and we incorporate this perspective in drawing inferences based on the characteristics exhibited by our agents. Thin slices of expressive behaviour are commonly used in zero acquaintance studies [45]. There is mixed evidence however as to whether slice length has an effect on accuracy. One might expect accuracy to increase in relation to increased exposure, i.e. increased accuracy as the amount of available information increases. A number of researchers report this is the case, e.g. Carney et al [47] reported increased accuracy for rating facial expressions based on observations ranging from 5s to 300s. Conversely, Ambady and Rosenthal's [43] meta-analysis found no linear increase in consensus correlations in slices between 30s and 300s. There seems to be little empirical agreement on whether length of video clip has any

effect on a rater's ability to detect personality [47]. However this may be due to the type of judgement being made. Some personality constructs have fewer or conflicting cues, and viewers may require the extra information provided by increasing exposure length, e.g. one of the most difficult personality traits to judge seems to be neuroticism [48]. As it would not have been wise at this stage of our research to consider the interaction between exposure length and personality type, we chose not to manipulate the length of exposure between agents, but to control observations at the minimum exposure reported for accuracy [43]. Length of clip was held at 30s for each agent.

The issue of location, i.e. *where* within a movie clip a slice should be extracted is more straightforward. Carney, Colvin and Hall [47, p.1059] reason that 'when strangers get to know each other' during real time dialogue, information contained early in the interaction may be less useful for making accurate personality assessments than the type of information available once the protagonists begin to relax, and get into conversation. These authors found accuracy was enhanced when ratings were based on later segments of a social interaction. Funder [49] similarly points to increased accuracy when 'good' information is taken from contexts where individuals can freely express their behavioural characteristics, and therefore provide insight as to their underlying personality traits. Such reasoning is fine within an Individual Differences research context. However the whole objective of our study was to get a bona fide assessment of the personalities currently depicted by our characters. Any attempt to proactively use stimuli containing 'good' information would conceivably be counterproductive and could arguably overestimate mean ratings. We thus opted for as 'natural' a scenario as possible, and chose an excerpt for each agent taken from the beginning of its interaction with a human user, i.e. what one would expect to see in the initial stages of communication when two strangers first meet up.

4.2 Method

Participants. 187 psychology students recruited from a university in Northern Ireland acted as raters (40 males, 147 females, ages ranging 18 to 47, mean age = 20.64). Evaluations of agent personality were completed over a period of five days, with the full sample split into five roughly equivalent groups based on individuals' scheduled laboratory sessions.

Materials and procedure. Three groups of raters (N=110) initially assessed each agent's personality based on still images of the agents' appearance alone, prior to viewing the video clips. The other two groups (N=77) viewed the video clips first, followed by the static images. For both static images and film clips, the order of presentation was also counterbalanced for each agent: for example, of the two groups who viewed the film clips first, one group observed Poppy, followed by Spike, followed by Prudence, and finally Obadiah. The other group viewed Obadiah first, followed by Prudence, followed by Spike, and lastly Poppy. In each sequence, personalities appear by alternating valence.

The static image display for each agent consisted of 3 coloured pictures of the agent (full face and three-quarter right and left profiles) displayed simultaneously on a screen via a ceiling-mounted AV projector. External conditions in terms of

lighting and camera angle are identical for each character.The image remained on screen while raters evaluated the agent in their own time (on average 2 minutes) using the six extraversion, six neuroticism and six psychoticism items taken from the abbreviated form of the Eysenck Personality Questionnaire Revised (EPQR-A) [50]. This is a forced-choice questionnaire with items rated as 'yes' or 'no'. The questionnaire doesn't test for the specific behaviours implemented, but for the general behaviours associated with the intended personality traits. Participants were asked to complete each of the eighteen items for each agent to rate (for example) 'how you think each statement describes Poppy'**.** We used the EPQR-A because of its brevity and because it has been shown to be a reliable and valid measure of Eysenck's three-factor model.

The 30sec video clip of each agent was also displayed on-screen via the ceiling-mounted AV projector. Characteristic of initial interactions between strangers, each clip featured eachagent's prototypical visual and vocal interaction with the same human user (N.I. male). This commenced with the agent introducing themselves and then inquiring how the user is feeling, what they would like to talk about etc. Examples of the capabilities of the architecture and system may be accessed via the SEMAINE website [51].

In order to focus raters' attention on the character itself, viewers were not shown visual images of the users, and users' vocal responses were visually (and not audibly) displayed in an instant messenger type format at the bottom of the screen. After viewing an agent's 30s video clip, raters again evaluated that agent using the EPQR-A. This process was reversed for the two groups who viewed the moving images first.

All raters then completed a five point Likert type adjective-based scale designed specifically for this study to provide additional indicators of agents' state. The scale consisted of fourteen adjectival descriptors constituting seven bipolar opposites: agreeable/disagreeable, interested/not interested, positive/negative, involved/indifferent, spontaneous/faked, sincere/not sincere, warm/cold. Participants finally rated characters' credibility in terms of familiarity, believability and consistency on a similar fivepoint Likert scale anchored at one end by the text 'very much' scoring 5, and at the other end by 'not at all' scoring 1.

4.3 Results

We initially assessed whether there were significant differences in mean overall EPQR-A ratings for each agent on each of the personality traits. Repeated measures ANOVAs show significant main effects of agent for extraversion ratings $F(3, 555) = 351.256$, $p<.001$; neuroticism $F(3, 558) = 166.055$, $p<.001$; and psychoticism $F(3, 555) = 476.703$, $p<.001$. Pair-wise comparisons reveal significant differences in scores (all at $p<.001$) between each agent and the respective other two for all 3 constructs judged. Poppy was rated significantly higher than the other agents for extraversion; Spike was rated significantly higher than other agents for psychoticism; Prudence was rated significantly lower than other agents for psychoticism; and Obadiah was rated significantly higher than other agents for neuroticism. Mean construct ratings for each agent are de-

tailed in Table 2.

We also explored the profile of traits (extraversion, neuroticism and psychoticism ratings) within each of the agents, again based on overall scores. Repeated measures ANOVAs demonstrate significant main effects of construct for each character: Poppy $F(2, 370) = 235.215$, $p<.001$; Spike $F(2, 372) = 310.704$, $p<.001$; Prudence $F(2,370) = 119.847$, $p<.001$; and Obadiah $F(2, 370) = 619.756$, $p<.001$. Pair-wise comparisons show significant trait score differences within each of the agents. Poppy's extraversion score is significantly higher than her psychoticism score ($p<.001$), which is in turn significantly higher than her neuroticism score ($p=.004$). Spike's psychoticism score is significantly higher than his neuroticism score ($p<.001$) which is significantly higher than his extraversion score ($p<.001$). Prudence's psychoticism score is significantly lower than her extraversion score ($p<.001$) which is significantly lower than her neuroticism score ($p<.001$). Obadiah's neuroticism score is significantly higher than his psychoticism score ($p<.001$) which is significantly higher than his extraversion score ($p<.001$).

Stimuli effects. In order to determine whether moving from picture to clip provided additional evidence in assessing personality, repeated measures analysis included order of rating as a factor. There were significant interactions between mode of stimuli and order of rating for the defined personality trait of all four agents: Poppy [$F(1, 184)=14.733$, $p<.001$, eta sq=.074]; Spike [$F(1,183)=6.520$, $p=.011$, eta sq=.034]; Prudence $F(1, 183)=8.607$, $p=.004$, eta sq=.045]; and Obadiah [$F(1, 184)=10.743$, $p=.001$, eta sq=.055]. Mean ratings and standard deviations for the intended trait for each agent are presented in Table 3. When the film clips were presented following the still images, mean trait ratings based on the clips for Poppy, Spike and Obadiah were significantly higher than mean ratings for still pictures. Mean trait ratings for Prudence are significantly lower as would be expected for this agent. Conversely when moving from clips to pictures there are no significant differences in mean ratings for Spike, Prudence and Obadiah. Poppy's extraversion score based on the film clip remains significantly higher than the picture rating for this agent. Graphical illustrations of these results are presented in Figure 2.

Text analyses. We further explored agents' ratings with respect to our adjectival descriptions, and report mean scores on the seven bipolar descriptors of overall levels of agreeableness, warmth, involvement, positivity, sincerity, interest, and spontaneity. Table 4details agent ratings (rank ordered) on the adjectival analyses. There were significant effects of agent for each description. Poppy's mean ratings are highest for six of the seven descriptors - agreeableness, interest, positivity, involvement, spontaneity, sincerity and warmth. Prudence scored highest on sincerity. Spike received the lowest mean scores on agreeableness, interest, warmth, and sincerity. Obadiah was rated lowest on positivity, involvement and spontaneity.

Credibility. Agents' familiarity, believability and consistency scores were reasonably high, all within the top half of the range. The overall mean rating for familiarity was 3.283. There was a significant effect of agent on mean familiarity ratings $F(3, 534) = 6.491$, $p<.001$. Rank ordered mean ratings for agents are illustrated in Table 5. Pairwise comparisons show Poppy is rated as significantly more familiar than all other agents (Spike $p<.001$, Prudence $p=.004$, and Obadiah $p=.003$). Spike's familiarity ratings are significantly lower than those of Prudence ($p=.043$).

Overall mean believability rating was slightly higher at 3.738. There was no significant main effect of agent on mean believability scores $F(3, 534) = 2.165$, $p=.091$. Poppy's mean ratings are lowest overall, and significantly lower than those of Obadiah ($p=.029$). There were no other significant differences in believability ratings.

Consistency ratings were highest, with an overall mean rating of 3.91. There was a significant effect of agent on consistency scores $F(3, 531) = 6.339$, $p<.001$. Poppy's mean ratings are significantly higher than those of Spike ($p<.001$) and Prudence ($p=.013$), and although higher than Obadiah's ratings, this is not significant. Spike's consistency rating is lowest, and significantly lower than the rating for Obadiah ($p=.001$).

4.4 Discussion

Our results provide what we believe is the first reported evidence of the viability of usage of human psychology constructs to define/shape behaviorof virtual characters. Our data indicate reliable signs of Big Three personality dimensions in animated characters as perceived by human viewers. As predicted, Poppy's overall mean extraversion rating was significantly higher than all other agents; mean psychoticism rating for Spike was significantly higher than other agents; mean psychoticism score for Prudence was significantly lower than ratings for other agents; and Obadiah's mean neuroticism rating was significantly higher than any other agent's rating. Furthermore, the significant trait differences revealed *within* each agent suggests each is perceived as a fundamentally distinctive character possessing a unique pattern of personality traits.

These overall findings are encouraging at this stage in the development of the agents, particularly in light of the differential effects of personality type on accuracy ratings reported in personality literature. Previous research points to differences in the accurate perception of personality depending on the construct being observed. Extraversion is considered to have the most observable cues and this trait has previously been accurately assessed with only minimal observation by strangers [52]. Psychoticism has fewer available cues, but this trait can still be relatively accurately assessed as the oppositeof agreeableness and conscientiousness (see [53] for agreeableness, and [52] for conscientiousness). On the other hand, neuroticism is one of the most difficult traits to detect, probably due to the lower number or conflicting nature of the behavioural cues reported in respect of this construct [54]. The fact that we have been able to bestow on Obadiah a set of behavioural cues which have been interpreted as representative of neuroticism is particularly encouraging. Furthermore,

these characteristics contributed to Obadiah being perceived as the 'least unbelievable' character, and to his behaviour being rated as the second most consistent.

That said, it maybe the case that our raters simply picked up on Obadiah's negative affect behaviour. The literature suggests that affect can be accurately judged during short observations [55] and this is consistent with an evolutionary approach to emotion [56]. Raters may have perceived Obadiah exhibiting generalized negative emotionality as opposed to the specific personality trait of neuroticism, and this might explain why this construct was so readily identified. Nonetheless, the fact that this agent was judged second lowest in psychoticism (also a negative trait) suggests that raters were indeed able to pick up on this agent's neurotic tendencies.

Stimuli effects. Analyses of ratings based separately on clips and still photographs provide similar sets of results and corroborate the distinction between characters at both a behavioural and a visual level. We also considered whether agents' behaviour provided any additional information which would help inform raters' judgement of personality type. A key finding is that moving from still picture to film clip provides additional information on each agent's intended personality type. The presence of behavioural characteristics reinforces the personality profiles that are already emerging from the still pictures. For agents' designed traits, the provided behavioural cues led to increased 'accuracy' against the criterion, or intended construct.

Crucially however, our analyses also indicate that each agent's rating on the two other *unintended* personality traits seem to be higher when based on visual appearance alone. For example Poppy's extraversion score is higher when behavioural cues are available, yet her neuroticism and psychoticism scores are higher when based purely on facial appearance. Whilst it cannot be excluded that such sharpening of profile would also be achieved by adding a limited number of additional still images, we believe that the addition of the behavioural elements is an important part of the current formation of the SAL system.

Credibility. Poppy is rated the most familiar and most consistent character, yet is not judged to be the most believable. A possible explanation, albeit speculative, is that the sense of familiarity and behavioural consistency exuded by Poppy's character may stem from raters' acquaintance with the coherent characteristics of a typically extraverted and attractive female cartoon character. Poppy may symbolise this archetypal character, and thus essentially be perceived as a simulated (and less credible) product. Further questions are raised in respect of Spike. Spike is both the least consistent and least familiar character, yet is not perceived to be the least believable. This may be due to those individuals who score high on psychoticism tending to portray behaviour which is perceived as *fundamentally* inconsistent by others. Raters may also have had less social interaction with those scoring high in psychoticism in general and thus found this character as least

familiar. Alternatively it may be the case that addressing the issue of 'believability'specifically in this way by human psychology methods is not fruitful. Nonetheless, our data suggest these agents 'work', because throughout reasonably high ratings of overall familiarity, believability and consistency have been attained.

Text analyses. Evaluation of agents on the adjectival descriptions further demonstrates that humans recognize agents' communicative behaviour as manifesting the personality type intended. Each SAL agent is perceived as showing the back-channel signals which are compatible with its emotional traits. Poppy and Prudence are perceived as conveying signals which express positive communicative intentions (e.g. agreeableness, sincerity; Spike and Obadiah are perceived as conveying negative communicative intentions (e.g. low agreeableness, low positivity).

Issues. Our results suggest that the behaviour of the artificial constructs which we have created seems capable of being interpreted as indicative of personality. These findings support work on personality perception in humans, indicating that when viewers are faced with artificial characters they seem to make similar personality inferences based on visual appearance and behavioural characteristics. Whilst the question of 'believability' remains outstanding, of key importance is the degree of success we have achieved in mapping the connections between behavioural characteristics assigned to our agents and the personality types as prescribed.

This study offers some insight into the benefits from behavioural cues. Raters' perception of agents' combined behavioural characteristics and facial appearance has tended to provide additional information in comparison to visual appearance alone. Nonetheless, further studies are required to tease out the contribution of different modalities by evaluating stimuli comprised solely of vocal information, visual information, and combined audio/visual behavioural cues.

A further issue relates to 'slice length'. It may be that thin slice judgements are accurate precisely because they are snap judgements, made quickly and in the absence of other distractions. In a context such as this however, it could be argued that 'less is best', and that flaws in the characters may actually become more obvious as information is increased through longer periods of observation.

There is also evidence that personality assessment may reflect not only the behavioural information depicted by the target, but also 'shared stereotypes'which may influence personality perception [57]. Whilst stereotypes can lead to biased/incorrect judgements[58] the psychological literature suggests that particularly in zero acquaintance research, accuracy may in fact be mediated by a kernel of truth hypothesis [59] i.e. how valid the stereotype is. Despite the fact that there *is* a shared stereotype that females tend to be more emotional than males, in this study it is Obadiah who receives the highest Neuroticism score (indicating emotional instability) and not Poppy or Prudence as might be expected. This may imply that for this character at least, personality perception has not relied on shared stereotypical responses. When there is

an increase in valid information, variables which lead to shared stereotypes tend to become less important [60], and inclusion of longer observation periods would help preclude any such prejudice.

Finally, the present effort highlights the interdisciplinary boundary as reflected in the use of the human psychology term 'credibility' and the Human Computer Interaction term 'believability'. Future researchers should be aware that there is no absolute overlap between these terms.

Conclusion. In conclusion, this work has been an initial attempt to provide a baseline evaluation of our characters as they currently exist. Based on psychological theory, four different characters have been created, each portraying distinctive behavioural characteristics indicative of human personality types. The credibility of the characters however is relative, and behavioural comparisons have been made against other artificial constructs. As the characters develop, future research will focus on the process of factoring out the differing components of personality. Further work will explore the relative efficacy of the differing channels of communication contained within each clip. We also intend to assess perception of the agents based on increased exposure, and to conduct studies where appearance and behavioural cues are contradictory. Evaluation of an agent for example, possessing Obadiah's facial appearance but displaying Poppy's behavioural characteristics will allow us to tease apart the relative contributions of each of these stereotypical components. This will build on our earlier work which shows that users perceive differences in the behaviours of an agent even if its appearance is the same [61]. The aim is to enhance agents' ability to interact believably with human users. Our evaluations indicate that this may be achievable. However a true test will require full evaluation of the actual interactions themselves.

REFERENCES

[1] D.K.J. Heylen, A. Nijholt and M. Poel, 'Generating Nonverbal Signals for a Sensitive Artificial Listener' in Verbal and Nonverbal Communication Behaviours, A. Esposito, M. Faunder-Zanny, E. Keller and M. Marinaro (eds), Lecture Notes in Computer Science, volume 4775, Springer Verlag, Berlin, pp. 264-274, 2007.

[2] P. Rizzo, M. Veloso, M. Miceliand A. Cesta, Personality-driven Social Behaviors in Believable Agents. In *Proceedings of the AAAI Fall Symposium on Socially Intelligent Agents*, pp. 109–114. 1997.

[3] B. Reevesand C. Nass. The Media Equation: How People Treat Computers, Television and New Media Like Real People and Places. CSLI Publications, Stanford, CA. 1996.

[4] A. Ortony, On making Believable Enotional Agents Believable, In R. Trappl, P. PettaandS. Payr (Eds.), *Emotions in humans and artifacts*. Cambridge, MA: MIT Press. 2003.

[5] E. Bevacqua, E. de Sevin, C. Pelachaud, M. McRorie and I. Sneddon, 'Building Credible Agents: Behaviour Influenced by Personality and Emotional Traits',Kansei Engineering and Emotion Research (KEER), pp. 1071-1080. 2010.

[6] R. McCrae and P. Costa, Validation of the Five-factor Model of Personality across Instruments and Observers. *Journal of Personality and Social Psychology*, vol. 52, pp. 81–90, 1987.

[7] H. Eysenck, *The Measurement of Personality*. Lancaster: Medical and Technical Publishers, 1976.

[8] L. Goldberg. The Structure of Phenotypic Personality Traits. *American Psychologist*, vol. 84, pp. 26-34. 1993.

[9] A. Saggino. The Big Three or the Big Five? A Replication Study. Personality and Individual Differences, vol 28(5), pp .879-886. 2000.

[10] A. Arya, L. Jeffries, J. Ennis and S.J. DiPaola, Facial Actions as Visual Cues for Personality. *Computer Animation and Virtual Worlds*, vol. 17, pp. 1–12, 2006.

[11] M. Mancini and C. Pelachaud, Distinctiveness in Multimodal Behaviors. In L. Padgham, D.C. Parkes, J. Muller and S. Parsons, editors, *Proceedings of Conference on Autonomous Agents and Multi-Agent Systems* (AAMAS08), pp. 159-166, 2008.

[12] B. Hartmann, M. Mancini, S. Buisineand C. Pelachaud, Design and Evaluation of Expressive Gesture Synthesis for Embodied Conversational Agents. In *3rd International Joint Conference on Autonomous Agents & Multi-Agent Systems*, Utrecht, pp. 1095-1096, 2005.

[13] C. Pelachaud and M. Bilvi, Modelling Gaze Behavior for Conversational Agents. In IVA03 *International Working Conference on Intelligent Virtual Agents*, Lecture Notes in Computer Science, volume 2792, pages 15–17, Germany, 2003. Springer.

[14] M. Mancini and C. Pelachaud, Dynamic Behavior Qualifiers for Conversational Agents. In C. Pelachaud, J-C. Martin, E. Andre, G. Chollet, K. Karpouzis, and D. Pel, editors, *Proceedings of 7th International Conference on Intelligent Virtual Agents*, volume 4722 of Lecture Notes in Computer Science, pages 112–124, Paris, France, 2007. Springer.

[15] P. Borkenau, N. Mauer, R. Riemann, F.M. Spinath and A. Angleitner. Thin Slices of Behavior as Cues of Personality and Intelligence. *Journal of Personality and Social Psychology*, vol. 86, pp. 599-614. 2004.

[16] I.S. Penton-Voak, N. Pound, A.C. Little andD.I. Perrett, Personality Judgments from Natural and Composite Facial Images: More Evidence for a Kernel of Truth' in Social Perception. Social Cognition, vol. 24, pp. 607–640. 2006.

[17] P. Ekman, W. Friesen and J.C. Hager. The Facial Action Coding System. London: Weidenfeld& Nicolson. 2002.

[18] J. Tipples, Wide Eyes and an Open Mouth Enhance Facial Threat. *Cognition and Emotion*, vol. 21, pp. 535-557, 2007.

[19] N. Fink, N. Neave, J.T. Manning and K. Grammar, Facial Symmetry and the Big-Five Personality Factors. *Personality and Individual Differences*, vol. 39, pp523–529, 2005.

[20] J. Walline. High Eye-Q. Science, vol 320 no. 5879, p.993. DOI: 10.1126/science.320.5879.993d. 2008.

[21] D. Farabee, R. Nelson and R. Spence, Psychosocial Profiles of Criminal Justice-and Non-Criminal Justice-Referred Clients in Treatment. *Criminal Justice and Behaviour*, vol. 20, pp. 336–346, 1993.

[22] D. Hebb, Drives and the CNS (Conceptual Nervous System). *Psychological Review*, vol. 62, pp. 243– 259, 1955.

[23] B. La France, A. Heisel and M. Beatty, Is There Empirical Evidence for a Nonverbal Profile of Extraversion? A Meta-analysis and Critique of the Literature. *Communication Monographs*, vol. 71, pp. 28–48, 2004.

[24] A. Gill and J. Oberlander, Taking Care of the Linguistic Features of Extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pp. 363–368, 2002.

[25] J. McCroskey, A. Heisel and V. Richmond.Eysenck's Big Three and Communication Traits: Three Correlational Studies. *Communication Monographs*, vol. 68, pp. 360–386, 2001.

[26] B. Smith, B. Brown, W. Strong and A. Rencher, Effects of Speech Rate on Personality Perceptions. *Language and Speech*, vol. 18, pp. 145-152, 1975.

[27] W. Exline, Affect Relations and Mutual Gaze in Dyads. In S. Tomkins and C. Izard, editors, *Affect, Cognition and Personality*. Springer, New York. 1965.

[28] K. Pennebaker and L.A. King, Linguistic Styles: Language Use as an Individual Difference. *Journal of Personality and Social Psychology*, vol. 77, pp. 1296–1312, 1999.

[29] R.J. Larsen and T.K. Shackelford, Gaze Avoidance: Personality and Social Judgements of People Who Avoid Direct Face-to-Face Contact. *Personality and Individual Differences*, vol. 21, pp. 907–917, 1996.

[30] M. Schröder, M. The SEMAINE API: Towards a standards-based framework for building emotion-oriented systems. *Advances in Human-Machine Interaction*. DOI http://dx.doi.org/10.1155/2010/319406, Article ID 319406, pp. 1-22, 2010.

[31] N. Wardand W. Tsukahara, Prosodic Features which Cue Back-channel Responses in English and Japanese. *Journal of Pragmatics*, vol. 23, pp. 1177–1207, 2000.

[32] R.M. Maatman, J.Gratch, and S. Marsella, Natural Behavior of a Listening Agent. In *5th International Conference on Interactive Virtual Agents*. Kos, Greece, pp. 25-36, 2005.

[33] D. Heylen, E. Bevacqua, M. Tellier and C. Pelachaud. Searching for Prototypical Facial Feedback Signals. In Proceedings of 7th International Conference on Intelligent Virtual Agents IVA 2007, pages 147–153, Paris, France. 2007.

[34] E. Bevacqua, D. Heylen, M. Tellier, and C. Pelachaud. Facial feedback signals for ECAs. In *AISB'07 Annual convention, workshop "Mindful Environments"*, pages147–153, Newcastle upon Tyne, UK, April 2007.

[35] G. McKeown, M.F. Valstar, R. Cowie and M. Pantic. "The SEMAINE Corpus of Emotionally Coloured Character Interactions", Proc. IEEE Int'l Conf. Multimedia & Expo, Singapore, Singapore, 1079-1084, 2010.

[36] R.R. McCrae and P.T. Costa. Personality in Adulthood: A Five Factor Theory Perspective'. New York. Guilford Press. 2003.

[37] P. Borkenau, N. Mauer, R. Riemann, F. Spinath, and A. Angleitner, Thin Slices of Behaviour as Cues of Personality and Intelligence. *Journal of Personality and Social Psychology*, vol 86, no. 4, pp. 599-614. 2004.

[38] D.C. Funder, and S.G. West, Consensus, Self-other Agreement and Accuracy in Personality Judgement: An Introduction. *Journal of Personality*, vol. 61, pp. 457-476. 1993.

[39] T.D. Letzring, S.M. Wells and Funder, D.C. Information Quantity and Quality Affect the Realistic Accuracy of Personality Judgment. Journal of Personality and Social psycjology, vol 91(1), pp111-123. 2006.

[40] N. Ambady, F.J. Bernieri, and J.A. Richeson, Toward a History of Social Behaviour: Judgemental Accuracy from Thin Slices of the Behavioural Stream. In *Advances in Experimental Social Psychology*, vol. 32, pp. 201- 271. 2000.

[41] R.R. McCrae, S.V. Stone, P.J. Fagan, and P.T. Costa, Identifying Causes of Disagreement Between Self-reports and Spouse Ratings of Personality. *Journal of Personality*, vol. 66, no. 3, pp. 285-313. 1998.

[42] A.R. King, and A.N. Pate, Individual Differences in Judgemental Tendencies Derived from First Impressions. *Personality and Individual Differences*, vol. 33, no. 1, pp. 131-145. 2002

[43] N. Ambady, and R. Rosenthal, Thin Slices of Behaviour as Predictors of Interpersonal Consequences: A Meta-analysis. *Psychological Bulletin*, vol. 111(2), pp. 256-274. 1992.

[44] M. Shevlin, S. Walker, M.N.O. Davies, P. Banyardand C.A. Lewis, Can You Judge a Book by its Cover? Evidence of Self-Stranger Agreement on Personality at Zero Acquaintance. *Personality and Individual Differences*, vol. 35, pp. 1373-1383. 2003.

[45] R. Hassin, and Y. Trope, Facing Faces: Studies on the Cognitive Aspects of Physiognomy. *Journal ofPersonality and Social Psychology*, vol. 78, pp. 837-852. 2000.

[46] C. Toris and B.M. DePaulo. Effects of Actual Deception and Suspiciousness of Deception on Interpersonal Perceptions. Journal of Personality and Social Psychology, vol. 47, pp. 1063- 1073. 1984.

[47] D.R. Carney, C.R. Colvin, and J.A. Hall, A Thin Slice Perspective on the Accuracy of First Impressions. *Journal of Research in Personality*, vol. 41, pp. 1054-1072. 2007

[48] D.C. Funder, and K. Dobroth, Differences Between Traits: Properties Associated with Inter-judge Agreement. *Journal of Personality and Social Psychology*, vol. 52, pp. 409-418. 1987.

[49] D.C. Funder, Accuracy in Personality Judgement: Research and Theory Concerning an Obvious Question. In B.W. Roberts and R. Hogan eds. *Personality Psychology in the Workplace: Decade of Behaviour*, pages 121-140. Washington: American Psychological Association. 2001

[50] L.J. Francis, L.B. Brownand R. Philipchalk, The Development of an Abbreviated Form of the Revised Eysenck Personality Questionnaire, (EPQR-A) – Its Use Among Students in England, Canada, the USA and Australia. *Personality and Individual Differences*, vol. 13, no. 4, pp. 443-449. 1992

[51] McKeown, G.J. (2011, Apr). Chatting with a Virtual Agent: The SEMAINE Project Character Spike [Video file]. Retrieved from http://youtu.be/6KZc6e_EuCg

[52] R. Lippaand J.K. Dietz, The Relation of Gender, Personality and Intelligence to Judges' Accuracy in Judging Strangers' Personality from Brief Video Segments. *Journal of Nonverbal Behaviour*, vol. 24, pp. 25-43. 2000.

[53] R. Gifford, Mapping Nonverbal Behaviour on the Interpersonal Circle. *Journal of Personality and Social Psychology*, vol. 61, pp. 279-288. 1999.

[54] D.C. Funderand C.D. Sneed, Behavioural Manifestations of Personality: An Ecological Approach to Judgemental Accuracy. *Journal of Personality and Social Psychology*, vol. 64, pp. 479-490. 1993.

[55] D. Matsumoto, J. LeRoux, C. Wilson-Cohn, J. Raroque, K. Kooken, P. Ekman, et al, A New Test to Measure Emotion Recognition Ability: Matsumoto and Ekman's Japanese and Caucasian Brief Affect Recognition Test (JACBART). *Journal of Nonverbal Behaviour*, vol. 24, pp. 179-209. 2000.

[56] C.E. Izard, *The Psychology of Emotions*. New York: Plenum Press. 1991.

[57] I.V. Blair, C.J. Judd, M.S. Sadler and C. Jenkins, The role of Afrocentric Features in Person Perception: Judging by Features and Categories. *Journal of Personality & Social Psychology*, vol. 83(1), pp. 5-25. 2002.
[58] S. E. Holleran, M.R. Mehl and S. Levitt. Eavesdropping on Social Life: The Accuracy Ratings of Daily Behaviour from Thin Slices of Natural Conversations. *Journal of Research in Personality*, vol. 43, pp660-672, 2009.
[59] D.A. Kenny and T.V. West. Self-perception as interpersonal perception. In J. Wood, J. Holmes, & A. Tesser (Eds.), *Self and relationships* (pp. 119-138). New York: Psychology Press. 2008
[60] J. Krueger, and M. Rothbart, Use of Categorical and Individuating Information in Making Inferences about Personality. *Journal of Personality and Social Psychology*, vol. 55, pp. 187-195. 1988.
[61] E. de Sevin, S. Hyniewska and C. Pelachaud. "Real-time Listener Action Selection for ECAs according to Personality". In Proceedings of Intelligent Virtual Agents 2010, IVA'10, Philadelphia, USA. pp. 187-193, 2010.

Acknowledgements
The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 211486 (SEMAINE).

Table 1:General description of the baseline of each SAL character for the two modalities face and gesture. The 6 expressivity parameters are the frequency,speed, spatial volume, energy, fluidity, and repetitivity

	Preference	**Frequency**	**Speed**	**Spatial volume**	**Energy**	**Fluidity**	**Repetitivity**
Poppy face	high	high	medium	high	medium	medium	medium
gesture	high	high	high	high	medium	medium	high
Spike face		medium	medium	medium	high	medium	high
gesture		medium	medium	medium	high	low	high
Prudence face	medium	medium	medium	medium	medium	medium	medium
gesture	medium	medium	medium	medium	medium	medium	medium
Obadiah face	medium	medium	low	medium	medium	low	low
gesture	low	medium	low	medium	medium	low	low

Table 2: Agents' EPQR-A (minval =0; maxval = 6) mean construct ratings

	Extraversion ratings			Psychoticism ratings			Neuroticism ratings		
Agent	Overall	Clip	Still	Overall	Clip	Still	Overall	Clip	Still
Poppy	4.97	5.63	4.31	2.4	2.13	2.67	1.94	1.16	2.72
Spike	1.85	1.76	1.95	5.07	5.21	4.94	2.56	2.36	2.77
Prudence	2.62	2.89	2.36	0.81	0.80	0.82	3.23	2.86	3.61
Obadiah	0.45	0.22	0.69	1.72	1.66	1.78	4.69	4.83	4.55

Table 3: Mean EPQR-A (minval = 0; maxval = 6) ratings for intented trait for each agent

	Picture ➔ Clip				Clip ➔ Picture			
Agent	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Poppy (extraversion)	3.82	2.420	5.64***	0.958	5.61**	0.891	5.01	1.618
Spike: high (psychoticism)	4.69	1.450	5.18**	1.075	5.26	0.938	5.29	0.971
Prudence: low (psychoticism)	0.89	1.212	0.65*	0.975	1.01	1.089	0.72	0.974
Obadiah (neuroticism)	4.27	1.772	4.92**	1.334	4.70	1.193	4.96	1.322

➔denotes order of presentation
***p<.001
**p<.01
*p<.05

Table 4: Comparisons between agents indicating rank order of mean ratings (value range 1-5) for adjectival descriptions

Adjectival descriptions (bipolar)	Mean ratings High						Low	
agreeableness	Poppy	8.20 >	Prudence	7.79>	Obadiah	5.38 >	Spike	3.77
interest	Poppy	7.89	Prudence	7.39	Obadiah	4.42_a	Spike	4.40_a
positivity	Poppy	8.83	Prudence	7.65	Spike	3.64	Obadiah	3.34
involvement	Poppy	7.48	Prudence	7.06	Spike	4.86	Obadiah	4.49
spontaneity	Poppy	6.32_a	Spike	6.01_a	Prudence	5.00	Obadiah	5.12
sincerity	Prudence	7.17_a	Poppy	6.65_a	Obadiah	6.15	Spike	5.32
warmth	Poppy	8.27	Prudence	7.13	Obadiah	4.51	Spike	3.65

Note. Agents within a row sharing a common subscript are not significantly different at the p<.05 level
>greater than

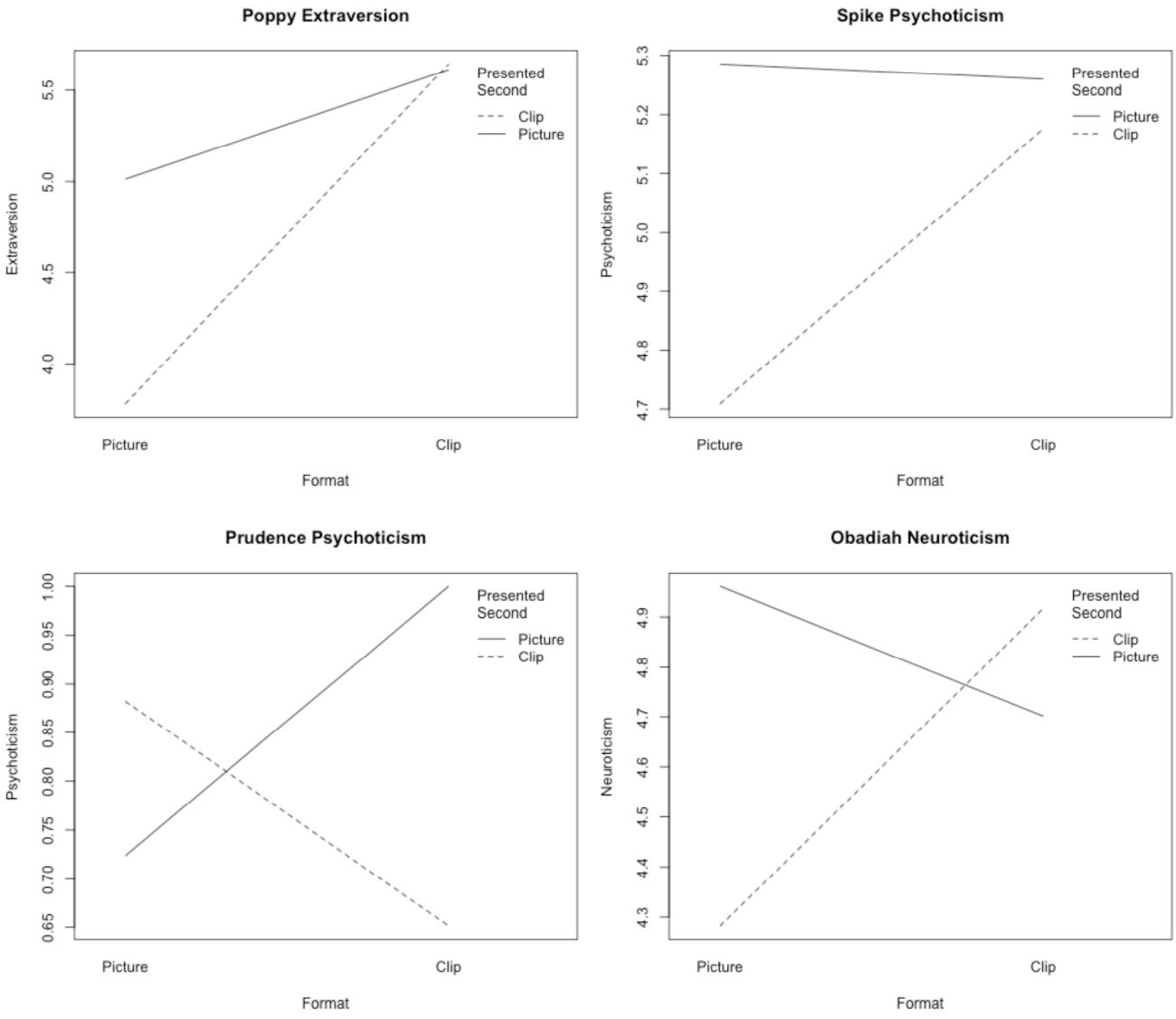
Table 5: Comparisons between agents indicating rank order of mean ratings (value range 1-5) for credibility

Credibility	Mean ratings High						Low	
Familiarity	Poppy	3.61>	Prudence	3.31>	Obadiah	3.14>	Spike	3.07
Believability	Obadiah	3.84	Spike	3.78	Prudence	3.70	Poppy	3.63
Consistency	Poppy	4.06	Obadiah	4.01	Prudence	3.85	Spike	3.70

>greater than

Fig 1. The 4 SAL agents (from left to right): Poppy, Spike, Obadiah and Prudence

Figure 2. Graphical illustration of mean EPQR-A (minval = 0; maxval = 6) ratings of agents' intended traits, based on order of rating


Poppy Extraversion
Extraversion
4.0
4.5
5.0
5.5
Picture
Clip
Format
Presented Second
Clip
Picture
Spike Psychoticism
Psychoticism
4.7
4.8
4.9
5.0
5.1
5.2
5.3
Picture
Clip
Format
Presented Second
Picture
Clip
Prudence Psychoticism
Psychoticism
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
Picture
Clip
Format
Presented Second
Picture
Clip
Obadiah Neuroticism
Neuroticism
4.3
4.4
4.5
4.6
4.7
4.8
4.9
Picture
Clip
Format
Presented Second
Clip
Picture

Margaret McRorie studied Psychology at Queen's University Belfast, graduating with a BA in 1997, followed by a PhD in 2001. She was appointed as Individual Differences lecturer in the Psychology Department at Queen's in 2002. Her research within the area of Individual Differences has continued, and she is currently focusing on individual differences issues within the area of emotion research. She is a certified FACS coder, and is specifically interested in exploring individual characteristics in the encoding and decoding of emotion. Her other research interests include emotion induction techniques, and cross-cultural differences in expressing and perceiving emotion. She is a member of the SEMAINE project.

Ian Sneddon graduated with a BSc in Psychology in 1974 followed by a PhD in Animal Social Behaviour, both from St Andrews University. He is currently a senior lecturer in the School of Psychology, Queen's University Belfast. Moving from ethological study of animal behaviour to the detailed observation of human behaviour, he is a certified FACS coder. Research interests include the expression and perception of emotion; emotion induction techniques and ethical issues; cross-cultural differences in emotion perception/expression. He was on the board of management of the HUMAINE network of excellence and is a member of the SEMAINE project.

Gary McKeown is a cognitive psychologist at the School of Psychology, Queen's University Belfast. His PhD explored mechanisms of implicit learning in the control of complex systems. His research focuses on communication, with interest in risk perception and decision making in environmental and health settings. This led to an interest in emotion and in particular the inter-relationship of cognition and emotion. Recent research has focused on emotion, social signal processing and the crosscultural emotion perception. He is a member of the SEMAINE project.

Elisabetta Bevacqua, post PhD at CNRS of Paris. She got her PhD in Computer Science at the University of Paris VIII in 2009. Since 2001, she works on Embodied Conversational Agents and her research field includes the verbal and non verbal communication, human-machine interaction and the implementation of models to simulate the humans' behaviour for virtual agents, particularly while listening to a user."

Etienne de Sevin after a master degree in cognitive sciences, received a PhD degree in Computer Science from VRLab EPFL, Switzerland in 2006. He currently works at LIP6 UMPC, Paris. His research interests focus on real-time action selection in autonomous virtual agents according to their internal (motivations, emotion, personalities) and their external (perceptions, social interactions) variables.

Catherine Pelachaud is Director of Research at CNRS in the laboratory LTCI, TELECOM ParisTech. She received her PhD in Computer Graphics at the University of Pennsylvania, Philadelphia, USA in 1991. Her research interests include embodied conversational agents, representation languages for agents, nonverbal communication, expressive behaviors and multimodal interfaces. She has been involved and is still involved in several national and European projects. She is an Associate Editor for the IEEE Transactions on Affective Computing.