AAAI Technical Report FS-12-02
Artificial Intelligence of Humor

Experimental Standards in Research on AI and Humor when Considering Psychology

Tracey Platt[1], Jennifer Hofmann[1], Willibald Ruch[1], Radoslaw Niewiadomski[2] & Jérôme Urbain[3]

University of Zurich, Binzmuehlestrasse 14/7, 8050 Zurich, Switzerland
Telecom ParisTech, Rue Dareau, 37-39, 75014 Paris, France
TCTS Lab, Univeristy of Mons, 31, Boulevard Dolez, 7000 Mons, Belgium

publications12@aaai.org

Abstract

Based on recent experiences between a laughing virtual agent and a human user at the intersection AI and humor and laughter, this paper aims to highlight some of the psychological considerations, when conducting AI and humor experiments. The systematic and standardized approach outlined in this paper will demonstrate how to reduce error variance that may be caused by confound variables such as having poor experimental controls. From the necessity of cover stories, protocols and procedures, the differences to the pros and cons of measuring subjectively and objectively and what is required so that both give valid and reliable results are offered as solutions to achieving this goal. Furthermore, the psychological individual differences that need consideration, such as the appreciation of different types of humor, mood, personality variables, for example, trait and state cheerfulness, and gelotophobia- the fear of being laughed at are discussed.

Introduction

The experimental study of humor stimuli in psychology started with Ertel (e.g., Ehrenstein and Ertel 1978) who varied structure and content in humor experimentally by forming sequences of words deviating from proper grammatical sequences and implementing taboo words (Ehrenstein and Ertel 1978). Others generated "artificial" humor stimuli and studied adjective-noun pairs varying in semantic distance (Godkewitsch 1974), computer-drawn caricatures with various degrees of exaggeration (Rhodes, Brennan and Carey 1987), or a weight-judging paradigm (Deckers 1993). These early approaches to artificial intelligence share several issues with contemporary computational approaches to humor. One is, that the humor is comparatively low in overall funniness compared to naturally occurring humor. As a consequence, refined measurement needs to be developed to be able to verify subtle differences at the lower end of the funniness spectrum. Another one is that, individual differences in the appreciation of different kinds of humorous stimuli will impact on the judgment of artificially created humor, independent of its "objective" quality.

Knowledge in psychology has grown over the past 100 years and sophisticated ways of standardized experimentation have been established. Humor research is occasionally experimental and a few issues could be adopted for the evaluation of artificial humor to increase the information outcome of evaluation studies. In the following some influential factors and structural requirements are highlighted and discussed. Also instruments are listed that might be adopted in empirical testing of artificial humor. Furthermore, we exemplify the topics discussed by giving examples from a recently conducted experiment on the impact of a laughing virtual agent on the subjective experience of humor when watching funny film clips.

A prototype of laughing virtual agent: a study at the intersection of AI, humor and laughter

Laughter is a significant feature of human communication, and machines acting in roles like companions or tutors should not be blind to it. So far, the progress has been limited that allows computer-based applications to deal with laughter and its recognition in the human user. In consequence, only few interactive multimodal systems exist that utilize laughter in interaction. Within the long-term aim of building a truly interactive machine able to

laugh and respond to human laughter, a prototype of laughing agent has been developed, basing on work of Urbain and colleagues (Urbain et al. 2010).

To evaluate this laughing virtual agent and its laughter, and experiment was designed to assess the impact of a virtual laughing companion on the humor experience of a human user. The experimental set up involved a participant watching a funny video with a virtual agent visually present on a separate screen. The expressive behavior of the virtual agent was varied among three conditions, systematically altering the degree of expressed appreciation of the clip (amusement) in verbal and non-verbal behavior, as well as different degrees of interaction with the participant's behavior. Furthermore, participant related variables were assessed with self-report instruments, which allowed for the investigation of the influence of mood and personality on the perception and evaluation of the virtual agent. Utilizing all these potential influences allowed for the control of systematic biases on the evaluation of the virtual agent, which is independent of its believability (e.g., individuals with a fear of being laughed at perceive all laughter negatively). The impact of the agent was assessed by investigating the influence of the session on participant's mood, as well as by self-report questionnaires assessing the perception of the virtual agent and the participant's cognitions, beliefs and emotions. Furthermore, the whole film-watching event was unobtrusively filmed, giving a close-up view of the participant's face, which allows for the coding of the facial responses. This study will serve as an example for the discussed experimental standards.

Psychological experiment considerations

Although there are many reference books that focus on experimental psychology, methods and standards in experiments, the topics discussed here have a special focus on conducting experiments in AI and humor.

The cover story

Curiosity, and social desirability often influence the way people will respond when they are asked to make a decision or judgment in psychological testing. This is especially true when these choices relate to humor, because a good sense of humor is often seen as a highly desirable personal quality (e.g., being good at judging the product of the experiment as amusing). Thus, participants will often over estimate the funniness of a joke, so as not to be seen as lacking a sense of humor. This can only happen if the participants can figure out beforehand what hypothesis the experiment is testing. One way of getting around this problem is to create a "false rationale" or "cover story". The cover story must have a credible context, which offers an explanation without alluding to the true nature of what is being measured. The cover story also moderates the expectations of the participant towards the ability of the system thus reducing any chance of priming them. For example, if the participant is told that so far no system was ever able to produce humorous puns before now, the participants might already rate slightly funny puns as medium funny, as they adapt to the fact that this is a huge technical improvement. Consequently, a good cover story gives the participant the idea that what is being measured does not relate to the generated puns, and then participants will be more likely to assess the puns at the true value without over-inflating them.

In respect to the example study, informing the participants beforehand as to what exactly the virtual agent was capable of doing would be essential. In the beginning of the experiment, the virtual agent should be set to perform some laughter. As the participant's responses to a humor stimulus were filmed, (with and without the virtual agent interaction) for later assessment, the elicited facial responses had to be derived only from the virtual agent interaction. This ensures, that the participant will not be surprised and amused solely by the fact the virtual agent can laugh, when eventually it does. The evaluation of the participant should go beyond the fact that the first laughing avatar exists.

If this information is not been made available to participants, it might be that the observed amusement is only due to the excitement/pleasure of the technical development of making a virtual agent laugh. This might be a research question in itself, but then, the comparison conditions should only entail non-laughing agents. Furthermore, the information to how much understanding the virtual agent has of the situation can be moderated: Participants will react differently when told that the agent has "no understanding at all" and just randomly displays laughter, or that the agent "understands humor and social interaction" and responds to the interaction with the participant. In the specific case, participants were told that they would be watching a video together with a virtual agent who would also be watching the film. Thus embedding the research in a specific cover story gives the participants something to focus on and gives the experimenter confidence that the required responses are being controlled as far as they can.

Standardization

Experimenters need to control the test environment sufficiently, to ensure that confounding variables do not influence the measurement and that any replication of the study would ascertain the same results. Therefore, it is necessary to create specific and pre-defined protocols on the procedure, including all verbal encounters with participants, a timeline including precise statements of how long different parts of the experiment last and where they take place, etc. These standards need to be followed

precisely for all aspects of the experiment. The clearer this protocol is, the easier it is for different experimenters to behave in the exact same way when engaging with participants. Thankfully, the very nature of AI facilitates standardized experimenting; as nothing apart from technical failure will deviate the program from what it was designed to do. However, if the AI has been programmed to create jokes for example, and the outcome or the success of the joke production is based on measuring a person's response to it, then ensuring that the environment the participant is in when they hear the joke allows to confidently state that the response (e.g., liking, funniness, laughter) was elicited only by the joke telling and not due to something else, like the presence of the experimenter. This is not as easy as it first sounds as the funnier someone finds a joke, the more susceptible they will be to finding the second joke even funnier, due to an increase of the positive affect-exhilaration (Ruch 1990).

Furthermore finding a satisfactory procedure that allows the experimental condition to be compared to a control condition is the essence of all experimental designs. During the example experiment participants were recruited through e-mail announcement of an "evaluation study of a laugh machine project". As an incentive, participants were offered a feedback on the questionnaire measures on request. It was announced that the study consisted of the filling in of questionnaires (approximately 30-45 minutes) and a session of 30 minutes on two given days. No further information on the aims of the study was given. Participants chose a date for the experimental session via the Internet and received confirmation by email.

At the experimental session, participants were welcomed by one of the two female experimenters and asked to fill in the STCI trait measure and the PhoPhiKat-45. Then, participants were asked to fill in the STCI-S to assess their current mood. Meanwhile, the participants were assigned to one of the three conditions. Afterwards, the second female experimenter accompanied the participant to the experimenting room, where the participant was asked to sit in front of a television screen. No informal talking that deviated from a pre-defined introduction script was adhered to in order that any mood state that had been noted in the STCI-S could be preserved until the actual experiment.

A camera allowed for the frontal filming of the head and shoulder, as well as upper body of the participant. Two male experimenters concerned with the technical components were present. Participants were asked for consent to have their shoulder and body movements recorded. They were also given headphones to hear the virtual agent. The experimenter explained that the participant was asked to watch a film together with the virtual agent and that the experimenters would always leave the room before the experiment started. Once the experimenters left the room, the virtual agent did greet the participant and subsequently, the video of the humor stimulus and the interaction with the virtual agent started. After the film, the experimenters entered the room again and the female experimenter accompanied the participant back to the location where the post measure of the STCI-S, as well as the evaluation questions for the session were filled in. After all questionnaires were completed, the first female experimenter debriefed the participant and asked for written permission to use the obtained data.

One consistent part of any experiment should be the experimenter. Therefore, the behavior of the experimenter should also be regulated in order that the experimenter/ participant interaction is not directly influencing anything that could alter the way in which the participant later engages with the virtual agent or humor stimuli of the AI system. For example males and females often interact to each other differently with laughter (e.g., Grammer and Eibl-Eibesfeldt 1990), there may also be an attraction component that may encourage someone being chatty or flirtatious, which may elevate or influence mood states that may increase noise in the data. By being consistent throughout all of the experiment- using the same experimenters, who act and say the same things, in the same way, to all participants will reduces error variance.

When designing an experiment, one should be aware of the influence the experimenters have and decisions as to how much interaction there should be, should be made. This is especially the case when covert filming will take place, as it is known that people are emotionally expressive and give a Duchenne smiles at the experimenter to signal that they are co-operating by conforming to what has been asked of them (Schug et al. 2010). So it may be that in such cases an intercom system be used rather than face-to-face to deliver instructions.

Subjective evaluation: Judging the stimulus vs. independent source of evidence

The appreciation, the understanding and the cognitions towards humor produced by AI can be assessed by self-report questionnaires. While those allow for an economical assessment of a participant's thoughts, feelings, and beliefs, they also entail certain biases. For example, social desirability (e.g., the wish to comply with societal rules or the experimenter) may influence a participant's rating. A further consideration relates to the terminology used in a questionnaire, which might be interpreted differently by participants (e.g., the word "natural" depends on whether natural in comparison to "human" or "natural" in comparison to the best possible virtual agent or pun generating computer). Also, answering scales might entail ambiguity: the term "often" might refer to different subjective values. Therefore, precautions should be taken to minimize those effects (e.g., label each step in

answering scales with words or percentages, provide brief definitions of words or items, give example items). Also, it should be considered which approach is used to create evaluation forms. Scales designed on principles of test construction theories with reports on quality criteria (objectivity, reliability, validity) are to be preferred. One might not need to design a new scale, if others have already put efforts in creating reliable measures.

To evaluate the quality of the interaction with the virtual agent, no comprehensive questionnaire existed which would assess the naturalness of the virtual agent and emotions, cognitions and beliefs toward it. As broad domains of human experience and thinking should be assessed, as well as a judgment of the system and the occurrence of social presence, a questionnaire was designed for the purposes of the experiment. The aim of the Avatar Interaction Evaluation Form (AIEF; Hofmann, Platt and Ruch, 2012) is to assess the perception of the agent, the emotions experienced in the interaction, as well as opinions and cognitions towards it on broad dimensions. The instrument consists of 32 items and 3 open questions, which were developed following a rational construction approach. The first seven statements refer to general opinions/beliefs and feelings on agents (e.g., "generally I enjoy interacting with avatars"). Then, 25 statements are listed to evaluate the experimental session. Following components are included: positive emotional experience, social presence aspects, and judgment of technical features of the avatar/believability, cognitive aspects assigned current agent. All statements are judged on a seven point Likert-scale (1 = "strongly disagree" to 7 = "strongly agree"). In the three open questions, participants can express any other thoughts, feelings or opinions they would like to mention, as well as describing what they liked best/least.

Facial expression: Objective supplement vs. independent sources of evidence

One resolution to occurring biases in ratings is to obtain objective measurements. Objective measurement methods allow for the assessment of an individuals' behavior (and also feelings) without relying on the participants self report. For example, there is agreement that seven (basic) emotions have a prototypical facial display, which is universal (if not regulated by social display rules). Therefore, experiences of joy, anger, fear, surprise, sadness, contempt and disgust (and more are discussed, cf. Keltner 1995) can be objectively assessed by analyzing a participant's facial expressions especially if the participant feels unobserved and does not voluntarily regulate their expressions. When considering emotional responses towards computer-generated humor or the responses towards a laughing virtual agent, it can be fruitful to film participants while they interact with the system or work on computer-based task and consequently analyze the spontaneous expression of emotion.

In respect to humor, amusement or exhilaration are facets of joy most strongly aligned with laughter and also smiling. In research on variations of smiling, different authors distinguish 14 to 18 different qualities of smiles (e.g., Ekman 1985). Of all these smiles however, only one smile is a signal for joy. The emotion of joy is accompanied by a facial configuration named by Ekman, Davidson, and Friesen (1990) as the *Duchenne smile* (to honor Duchenne, who first described how this pattern distinguished enjoyment smiles from other kinds of smiling) but also referred to as a felt smile, enjoyment smile or Duchenne display (Ekman et al. 1990, Frank and Ekman 1993). The Duchenne smile is defined by a simultaneous and symmetric contraction of zygomatic major muscle and orbicularis oculi pars orbitalis muscle and it differs from other smiles on the basis of timing, and other factors. Differences between Duchenne and non-Duchenne smiles were found in intensity (e.g., Krumhuber and Manstead 2009), symmetry (e.g., Ekman et al. 1981; Hager and Ekman 1997) and dynamics (e.g., Ambadar et al. 2009; Krumhuber and Kappas 2005).

Considering the relationship of smiling and laughter, Ruch (1990, 1993) found laughter occurring in response to humorous stimuli and generally joy, suggesting a link to the DD (Ruch 1993). Consequently, he argues that the difference between smiling and laughter may be a difference in intensity of the emotion amusement. Different authors (Keltner and Bonanno 1997; Ruch 1993) defined the basis of joyful/amused laughter (Duchenne laughter) to consist of the DD plus an audible, laughter-related vocalization and open mouth. Duchenne laughter is typically lasting longer than Duchenne smiling, entails a more intense contraction of the zygomatic major muscle (Ruch and Ekman 2001) and compared to smiling people, laughing people report higher perceived funniness of jokes (e.g., Ruch 1990).

Therefore, the assessment of Duchenne smiles and laughs may indicate amusement induced by computer-generated humor. Furthermore, non-emotion relevant facial behaviors can be equally significant and provide information. Coming back to the example of watching a funny movie along with a virtual agent, it might be indicative to assess how may times the participant actually looked at the avatar, instead of focusing on the screen with the funny movie. Measuring the eye movement (gaze behavior) could also be used. The amount of attention drawn by the avatar might be an indicator for the social presence felt, or the connection built between the participant and the virtual social partner.

Thus, research has shown that this facial marker, linked to enjoyment and therefore useful in the evaluation of AI and humor, carries potentially valuable stand-alone, or

supplementary, objective information. Moreover, assessment tools are available in the form of facial electromyography (or facial EMG) or coding systems that can assist in collecting this data. The former is economic but may also be less valid, for a number of reasons. For example, a surface (or needle) electrode gathers the signals generated by the muscles contracting. A reference electrode is also needed. This can be very restrictive and inhibit spontaneous humor responses. Furthermore, although the fine-grained intensity may vary, the amplitude is not standardized and so it cannot be compared across people. This is due to the strength of the signal is not only affected by degree of muscle contraction, but by factors, such as muscle thickness, exact muscle placement, skin thickness and the fat layers etc.

Coding systems differ in the level of sophistication and usually trade how much information they offer to how time consuming the coding process is. The Facial Action Coding System (FACS; Ekman, Friesen and Hager, 2002) is the leading coding scheme that offers a reliable, valid and objective assessment of all visually discernible facial action. The Facial Action Coding System (FACS; Ekman et al. 2002) is an anatomically based, comprehensive technique, which distinguishes 44 action units (AUs). These are the minimal units that are anatomically separate and visually distinguishable. FACS also allows for measurement of the timing of a facial movement, its symmetry and intensity, and its degree of irregularity of onset, apex or offset as well as several categories for head and eye positions/movements and miscellaneous actions. Using FACS and viewing digital-recorded facial behavior at frame rate and in slow motion certified FACS coders are able to distinguish and code all visually discernible facial expressions. Many studies and experiments on the sense of humor have applied the FACS to assess individual's emotional responses and have found moderate convergences to self-reports of funniness. Nevertheless, the objective assessment has been shown to be superior when assessing humorous traits, which may be influenced by social desirability.

Controlling error variance and understanding individual differences in humor: The role of mood and personality

Considering the individual differences of the humor preferences of participant's as well as personality traits and mood states are all necessary in order to maximize the amount of valid conclusions can be ascertained from the information gathered from the experiment.

Mood states as predictors

Like funniness and aversiveness, positive and negative affect, are orthogonal factors. The supertrait extraversion predicts individual differences in positive affect and neuroticism accounts for individual differences in negative affect. Can these relationships be found in the realm of humor appreciation as well?

There is, indeed, a consistent positive inter-correlation among appreciation of the three humor categories, which is low for funniness but relatively high for aversiveness. Thus, there is some room left for the assumption of stable individual differences in the tendencies to find humor generally more aversive or generally funnier. Since funniness represents the positive responses to humor and aversiveness covers the possible negative ones it could be hypothesized that extraversion correlates positively with funniness of the three humor categories and neuroticism predicts their aversiveness. However, in a review of studies, Ruch (1992) found only spurious effects of extraversion on generalized positive responses to humor. While the zero-order coefficients obtained were overwhelmingly in the expected direction, they generally lack both statistical and practical significance.

Personality as disposition to responses (trait cheerfulness)

Humor was claimed to involve a non-bona-fide mode of communication, and people need to process humor in a playful frame of mind. The assessment of personality variables allows for a control of habitual factors influencing the perception of AI generated humor or, in the example mentioned, the virtual agent, independent of its objective quality/believability. By assessing humor relevant traits (and states), individuals who for example misjudge humor or laughter can either be excluded from further analysis, or the influence of traits can be investigated for the dependent variables.

Ruch and colleagues conducted a series of studies based on the observation of interindividual and intraindividual differences in humor-related behavior. Certain individuals tend to generally appreciate, create, or laugh more easily and intensively at humorous stimuli than others do. Besides, there are actual dispositions for humor, varying across time and context. In Ruch's model, both state and trait cheerfulness, seriousness and bad mood are operationalized in facets. The model does not claim comprehensiveness for all humor-related behaviors, but while the expression of humor may be culture specific and changing over the course of time, the affective and mental foundations may be universal.

While habitually serious people will be less likely to process humor, people in a cheerful mood will be more ready to laugh and people with a cheerful trait will have a lowered threshold for smiling and lifting themselves into a cheerful state. Thus, is assumed that cheerfulness and seriousness (but also bad mood, as a further marker of humorlessness) as states and traits should play a role in understanding humor. A state-trait model of cheerfulness,

seriousness and bad mood was introduced to describe the temperamental basis of humor. Cheerfulness represents an individual's actual or habitual disposition for amusement, laughter and seeing the bright side of life. Trait seriousness and trait bad mood represent dispositions for different forms of humorlessness and lower the threshold for engaging in humor and displaying smiling and laughter, though for different reasons. It is expected that those traits and states moderate responses in experiments on AI and humor and experimenter should include those traits and states to separate the effects of personality on humor appreciation and engagement in humor, from task related appreciation and engagement. In the example study, it was assumed that individuals with high scores in trait cheerfulness and state cheerfulness would more easily engaged with the virtual avatar and would be more easily influenced by its laughter, independent of the condition. On the other hand, individuals with high scores in state bad mood should have enhanced thresholds for being amused and respond less amused towards the agent. In respect to state seriousness, individuals with a high score will focus on the task and prefer a sober, object-oriented and rational style. Consequently, they may also be less likely to engage with the agent. Furthermore, the individual differences in traits may also influence the changes in mood throughout the experiment. For trait cheerful individuals, state cheerfulness should increase due to the humor intervention, while seriousness and bad mood should drop.

The State-Trait Cheerfulness Inventory (STCI; Ruch et al. 1996, 1997) is the instrument that assesses the temperamental basis of the sense of humor in the three constructs of cheerfulness (CH), seriousness (SE), and bad mood (BM) as both states (STCI-S) and traits (STCI-T). The standard trait form (STCI-T<60>; Ruch et al. 1996) is a 60-item self-report instrument. It applies a four-point answer format (1 = strongly disagree to 4 = strongly agree), to assessing the traits of cheerfulness, seriousness, and bad mood with 20 items for each scale. The standard state form (STCI-S<30>; Ruch et al. 1997) assesses the respective states of cheerfulness, seriousness and bad mood with ten items each. Ruch and Köhler (2007) report high internal consistencies for the traits (CH: .93, SE: .88, and BM: .94). The one-month test-retest stability was high for the traits (between .77 and .86), but low for the states (between .33 and .36), conforming the nature of enduring traits and transient states. The state version of the scale can be used to test people pre and post experiment to evaluate any arousal effect of cheerfulness over the course of the humor intervention.

Personality and dispositions to liking of different humor types (3WD)

From an individual difference perspective regarding the perception of humor, theorists have long recognized two sources, content and structure (or: joke work vs. tendency; thematic vs. schematic; cognitive vs. orectic factors). Utilizing a factor analytic approach confirmed that two components are strong sources of variance. The two structural factors consistently appearing are incongruity-resolution (INC-RES) humor and nonsense (NON) humor. Irrespective of the fact that the jokes and cartoons relating to these factors had different contents for example the themes or the targets of the joke, they always share structural properties and the way they are processed.

Punch lines that have a surprising incongruity, which can be completely resolved, characterize jokes and cartoons that pertain to INC-RES humor. The common component of this type of humor is that the recipient first discovers an incongruity, which is then fully resolvable upon consideration of information embodied in the joke or cartoon.

The surprising or incongruous punch line was also found for Nonsense humor, however, the punch line may provide no resolution or only a partial resolution (leaving an essential part of the incongruity unresolved), or create new absurdities or incongruities. In nonsense humor the resolution information gives the appearance of making sense out of incongruities without doing so. A third factor, sexual (SEX) humor, may have either structure, but is homogeneous with respect to the sexual content.

The 3 WD (3 Witz-Dimensionen) Test of Humor Appreciation is a performance test measuring both funniness and aversiveness of jokes and cartoons for the three humor categories, incongruity-resolution humor, nonsense humor, and sexual humor. The instrument contains 35 (forms A and B) jokes and cartoons, which are rated on "funniness" and "aversiveness" using two 7-point scales. The funniness rating ranges from not at all funny = 0 to very funny = 6. The aversiveness scale ranges between not at all aversive = 0 to very aversive = -6. The first five items are to "warm up" and therefore are not scored is presented in a test booklet containing two or three jokes or cartoons per page. The instructions are typed on separate answer sheets containing the two sets of rating scales.

Six scores can be derived from each form of the test: three for funniness of incongruity-resolution, nonsense and sexual humor (i.e., INC-RESf, NONf, and SEXf) and three for their aversiveness (i.e., INC-RESa, NONa, and SEXa). The six scores generated indicate an individual's humor preference at a general level. Structure preference index SPI (obtained by subtracting INC-RES from NON, indicating the relative preference for one type over the other) and an index of liking of sexual content (build by removing the variance due to liking of content) have been derived and validated. Funniness and aversiveness of a humor type may be combined to form a general appreciation score.

Gelotophobia

One major concern when measuring the humor of people and their responses to humor stimuli, is that for some, they lack the perception of the enjoyment of laughter and even have a fear of being laughed at (gelotophobia). This lack of perception they experience will give rise to problems in any study involving the elicitation of humor. Gelotophobes assume all the laughter is directed at them in a negative, malicious, way. They seem to perceive all good-natured humor as negative derision (Platt 2008, Ruch, Alfreder and Proyer 2009). Although some gelotophobes can create humor in a study investigating the humor of gelotophobes Ruch, Beermann and Proyer (2009) found that gelotophobes are less cheerful and evaluate their own humor style as inept. Consistently, gelotophobes over-inflate responses that relate to the lower propensity they appear to have for joy, happiness, amusement or more specifically to humor appreciation. Therefore, the laughing agent might be interpreted as a threat and the evaluation would be biased by the individuals fear. By assessing the gelotophobic trait, individuals with at least a slight fear of being laughed at can either be excluded from further analysis, or the influence of gelotophobia can be investigated for the dependent variables.

The construct of gelotophobia was primarily observed in interaction between therapist and patient. An article by Dr. Titze (2009) described the assessment in a clinical setting. The descriptive criteria provided by Titze allowed Ruch and Proyer (2008a; 2008b) to build an effective and efficient 15-item self report instrument that allowed to identify gelotophobes at a non, slight, pronounced and extreme level of the this individual difference trait. The main instrument for the assessment of gelotophobia is the GELOPH <15> (Ruch and Proyer 2008b), which contains 15 statements reflecting the phenomenological world of gelotophobes. A hierarchical factor analysis (Platt, Ruch, Hofmann and Proyer 2012) showed that these components of gelotophobia load onto three clear factors, the first factor being controlling ones environment in order to avoid laughter by withdrawal or internalizing that one is a valid object of derision, the second factor being a paranoid sensitivity to anticipated ridicule and the third was having a disproportionately negative response to being laughed at.

Further, the joy of being laughed at (gelotophilia) and the joy of laughing at others (katagelasticism) might alter the experience with the agent, as katagelasticists might enjoy laughing at the agent, while gelotophiles may feel laughed at by the agent and derive pleasures from this. Both dispositions may increase the positive experience of interacting with an agent. The PhoPhiKat-45 is a 45-item measure of gelotophobia ("When they laugh in my presence I get suspicious"), gelotophilia ("When I am with other people, I enjoy making jokes at my own expense to make the others laugh"), and katagelasticism ("I enjoy exposing others and I am happy when they get laughed at"). Answers are given on a 4-point Likert scale (1 = strongly disagree to 4 = strongly agree). Ruch and Proyer (2009) found high internal consistencies (all alphas ≥ .84) and high retest-reliabilities ≥ .77 and ≥ .73 (three to six months). In the present sample, reliabilities were satisfactory to high and ranged between α = .81 to .83.

Conclusion

Bringing together the fields of AI and psychology can only strengthen both. Understanding the psychological impact of the interface between computer and human allows for the evaluation of the AI's success. However, safeguards must be taken in order to benefit from extrapolating the information gathered. This is especially so when evaluating humor, as many aspects of humor revolve around the individual differences in personal preferences in the type of humor the individual finds funny and also personality traits. It is not only these stable factors that play a role, the mood of the participants, at the time of the experiment, will also influence their answer style. This mood needs to be measured before and after the AI interaction, which in tern needs to be preserved by the experimenter, in order to be taken into account as part of the assessing the humorousness of the AI. By taking the time to formalize the experiments and experimental environment as much as possible, by standardizing and adhering to protocols, developed for and used in psychological testing, one can ensure less error variance and more valid and reliable conclusions to be made.

Acknowledgement

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n°270780 (ILHAIRE project).

References

Ambadar, Z., Cohn, J. F., & Reed, L. I. 2009. All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. *Journal of Nonverbal Behavior, 33,* 17-34.

Deckers, L.H. 1993. On the validity of a weight-judging paradigm for the study of humor. *Humor, 6,* 43 - 56.

Ehrenstein, W. H., & Ertel, S. 1978. Zur Genses des Lustigkeitseindrucks. *Psychologische Beiträge, 20,* 360 - 374.

Ekman, P. 1985. *Telling lies. Clues to deceit in the marketplace, politics, and marriage.* New York: W.W. Norton.

Ekman, P. 1999. Basic emotions. In T. Dalgleish & M. Power (Eds.), *Handbook of cognition and emotion* (pp. 45 - 60). Sussex: John Wiley & Sons Ltd.

Ekman, P. & Friesen, W. V. 1971. Constants across cultures in the face and emotion. *Journal of Nonverbal Behavior, 6,* 238 - 252.

Ekman, P. & Friesen, W. V. 1982. Felt, false and miserable smiles. *Journal of Personality and Social Psychology, 17,* 124-129.

Ekman, P., Davidson, R. J., & Friesen, W. V. 1990. The Duchenne smile: Emotional expression and brain physiology II. *Journal of Personality and Social Psychology, 58,* 342-353.

Ekman, P., Friesen, W. V., & Hager, J. C. 1981. The symmetry of emotional and deliberate actions. *Psychophysiology, 18,* 101-106.

Ekman, P., Friesen, W. V., & Hager, J. C. 2002. *Facial Action Coding System: A technique for the measurement of facial movement.* Palo Alto: Consulting Psychologists Press.

Frank, M. G. & Ekman, P. 1993. Not all smiles are created equal: The differences between enjoyment and non-enjoyment smiles. *Humor, 6,* 9-26.

Godkewitsch, M. 1974. Correlates of humor: Verbal and nonverbal aesthetic reactions as functions of semantic distance within adjective-noun pairs. In D. E. Berlyne (Ed.), *Studies in the new experimental aesthetics* (pp. 279 - 304). Washington, DC: Hemisphere.

Grammer, K., & Eibl-Eibesfeldt, I. 1990. The ritualization of laughter. In W. A. Koch (Ed). *Natürlichkeit der Sprache und der Kultur: acta colloquii* (pp. 192 - 214). Bochum: Brockmeyer.

Hager, J. C., & Ekman P. 1997. The asymmetry of facial actions is inconsistent with models of hemispheric specialization. In: P. Ekman & E. Rosenberg (Eds.). *What the face reveals* (pp. 40 - 62). New York: Oxford University Press.

Hofmann, J., Platt, T., & Ruch, W. 2012. *Avatar Interaction Evaluation Form (AIEF).* Unpublished research instrument, University of Zurich.

Keltner, D. 1995. Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality and Social Psychology, 68,* 441- 454.

Keltner, D., & Bonanno, G. A. 1997. A study of laughter and dissociations: distinct correlates of laughter and smiling during bereavement. *Journal of Personality and Social Psychology, 73,* 687-702.

Krumhuber, E. G., & Kappas, A. 2005. Moving smiles: the role of dynamic components for the perception of the genuineness of smiles. *Journal of Nonverbal Behavior, 29,* 3 – 24.

Platt, T. 2008. Emotional responses to ridicule and teasing: Should gelotophobes react differently? *Humor, 21*, 105-128.

Rhodes, G., Brennan, S., & Carey, S. 1987. Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive Psychology, 19,* 473-497.

Ruch, W. 1990. *Die Emotion Erheiterung: Ausdrucksformen und Bedingungen* [The emotion of exhilaration: Forms of expression and conditions.] Unpublished habilitation thesis, University of Düsseldorf, Germany.

Ruch, W. 1992. Assessment of appreciation of humor: Studies with the 3 WD humor test. In C. D. Spielberger and J. N. Butcher (Eds.), *Advances in personality assessment* (pp. 27- 75). New Jersey: Lawrence Erlbaum Associates.

Ruch, W. 1993. Exhilaration and humor. In M. Lewis & J. M. Haviland (Eds.), *The Handbook of Emotions* (pp. 605-616). New York: Guilford Publications.

Ruch, W., & Ekman, P. 2001. The expressive pattern of laughter. In A. W. Kaszniak (Ed.), *Emotion, qualia, and consciousness* (pp. 426-443). Tokyo: Word Scientific Publisher.

Ruch, W., Altfreder, O., & Proyer, R. T. 2009. How do gelotophobes interpret laughter in ambiguous situations? An experimental validation of the concept. *Humor, 22 (1/2),* 62-89.

Ruch, W., Beermann, U., & Proyer, R. T. 2009. Investigating the humor of gelotophobes: Does feeling ridiculous equal being humorless? *Humor, 22,* 111-143.

Ruch, W., & Köhler, G. 2007. A temperament approach to humor. In W. Ruch (Ed.), *The sense of humor: Explorations of a personality characteristic* (pp.203 - 228). Berlin: Mounton de Gruyter.

Ruch, W., Köhler, G., & van Thriel, C. 1996. Assessing the "humorous temperament": Construction of the facet and standard trait forms of the State-Trait Cheerfulness Inventory-STCI. *Humor, 9,* 303-339.

Ruch, W., Köhler, G., & van Thriel, C. 1997. To be in good or bad humour: Construction of the state form of the State-Trait-Cheerfulness-Inventory—STCI. *Personality and Individual Differences, 22*, 477-491.

Ruch, W., & Proyer, R. T. 2008a. The fear of being laughed at: Individual and group differences in gelotophobia. Humor: International *Journal of Humor Research, 21,* 47-67. doi:10.1515/HUMOR.2008.002

Ruch, W., & Proyer, R. T. 2008b. Who is gelotophobic? Assessment criteria for the fear of being laughed at. *Swiss Journal of Psychology, 67,* 19-27. doi:10.1024/1421-0185.67.1.19

Ruch, W., & Proyer, R. T. 2009. Extending the study of gelotophobia: On gelotophiles and katagelasticists. Humor: International *Journal of Humor Research, 22,* 183-212. doi:10.1515/HUMR.2009.009

Schug, J., Matsumoto, D., Horita, Y., Yamagishi, T., & Bonnet, K. 2010. Emotional expressivity as a signal of cooperation. *Evolution and Human Behavior, 31,* 87 – 94.

Titze, M. 2009. Gelotophobia: The fear of being laughed at. *Humor: International Journal of Humor Research, 21,* 27-48. doi:10.1515/HUMR.2009.002

Urbain, J., Niewiadomski, R., Bevacqua, E., Dutoit, T., Moinet, A., Pelachaud, C., Picart, B., Tilmanne, J., Wagner, J., 2010. AVLaughterCycle. Enabling a virtual agent to join in laughing with a conversational partner using a similarity-driven audiovisual laughter animation, *Journal of Multimodal User Interfaces, 4*, 47-58.