Virtually Bad: A Study on Virtual Agents that Physically Threaten Human Beings

Socially Interactive Agents Track

Tibor Bosse
Vrije Universiteit Amsterdam
Dept. of Computer Science
De Boelelaan 1081
1081 HV Amsterdam
The Netherlands
t.bosse@vu.nl

Tilo Hartmann
Vrije Universiteit Amsterdam
Dept. of Communication Science
De Boelelaan 1105
1081 HV Amsterdam
The Netherlands
t.hartmann@vu.nl

Romy A.M. Blankendaal
Vrije Universiteit Amsterdam
Dept. of Computer Science
De Boelelaan 1081
1081 HV Amsterdam
The Netherlands
r.a.m.blankendaal@vu.nl

Nienke Dokter
Vrije Universiteit Amsterdam
Dept. of Communication Science
De Boelelaan 1105
1081 HV Amsterdam
The Netherlands
n.dokter@live.nl

Marco Otte
Vrije Universiteit Amsterdam
Dept. of Computer Science
De Boelelaan 1081
1081 HV Amsterdam
The Netherlands
m.otte@vu.nl

Linford Goedschalk
Vrije Universiteit Amsterdam
Dept. of Computer Science
De Boelelaan 1081
1081 HV Amsterdam
The Netherlands
l.goedschalk@hotmail.com

ABSTRACT

This paper introduces the concept of "virtual bad guys": intelligent virtual agents that take a negative or even aggressive stance towards the user. Although they pave the way to various interesting applications, it is hard to create virtual bad guys that are taken seriously by the user, since they are typically unable to apply serious sanctions. To address this issue, this study experimentally investigated the effect of "consequential" agents that are able to physically threaten their human interlocutors. A consequential agent was developed by equipping users with a (non-functioning) device, through which they were made to believe the agent could mildly shock them. Effects on participants' levels of anxiety and (physiological and self-reported) stress were measured, and the role of presence and perceived believability of the virtual agent was assessed. The consequential agent triggered a stronger physiological stress response than the non-consequential agent, whereas self-reported levels of anxiety and stress did not significantly differ. Furthermore, while presence and believability were substantially associated with users' stress response, both states did not mediate or explain the effect of a consequential vs. non-consequential agent on stress, as they did not significantly differ between conditions. Implications of these findings and suggestions for follow-up studies on "virtual bad guys" are discussed.

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10-15, 2018, Stockholm, Sweden. © 2018, International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

KEYWORDS

Human-Agent Interaction; Virtual Reality; Aggression; Stress.

ACM Reference format:

T. Bosse, T. Hartmann, R.A.M. Blankendaal, N. Dokter, M. Otte, and L. Goedschalk. 2018. Virtually Bad: A Study on Virtual Agents that Physically Threaten Human Beings. In Proc. of the 17th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2018), July 10-15, 2018, Stockholm, Sweden, ACM, New York, NY, 9 pages.

1 INTRODUCTION

Intelligent Virtual Agents (IVAs) are intelligent digital interactive characters that can communicate with humans and other agents using natural human modalities like facial expressions, speech, gestures and movement [4]. Over the past decades, IVAs have become widely used in a variety of application domains, including education [22], healthcare [12], and the military [41]. In such domains, IVAs are typically incorporated into applications in which they interact with human users while playing a role that was traditionally played by human beings, such as teacher, therapist, or teammate.

Interestingly, in the vast majority of these cases, IVAs take a "positive" attitude towards the user. That is, they aim to support the user with a task or help to deal with a problem. Instead, the area of IVAs with a "negative" or aggressive attitude towards users (to which we refer as 'virtual bad guys' in this paper) has been heavily under-researched. This is a missed opportunity, as the concept of virtual bad guys opens up a range of useful applications, including virtual training of conflict resolution skills [7], virtual reality exposure therapy [35], and anti-bullying education [43].

However, *believability*, a common problem in the design of IVAs, poses a particular challenge for virtual bad guys. For IVAs, being believable can be defined as providing the illusion of being alive [3], and important requirements for achieving this include not only a realistic appearance, but also human-like functional and social qualities [10]. Unfortunately, this poses a particular challenge for virtual bad guys, because effective applications require that users feel seriously threatened or stressed by the IVA [32]. After all, interacting with an aggressive individual in real life always brings along the risk of being attacked. This is difficult to achieve for IVAs, since they are usually 'non-consequential', i.e., are unable to apply serious sanctions to users. As a result, users also perceive and categorize IVAs as virtual beings that have no influence in the real world [18]. These factors plausibly shape and skew how humans respond to IVAs with a negative attitude.

Accordingly, the study reported in this paper addresses the challenge of how to develop virtual bad guys that are taken seriously. Inspired by earlier research on aggressive virtual agents [6], we operationalize the notion of "being taken seriously" as "being able to induce anxiety and stress". Hence, the following research question is addressed in particular: "Does an IVA induce higher levels of anxiety and stress if it is able to physically threaten its users?". The general approach includes the design and experimental examination of the effects of a technologically advanced IVA that is able to physically threaten its users (which we will call *consequential interaction*). First, a virtual reality (VR) application was developed, involving a virtual bad guy that has the ability to interact with users through speech. Based on spoken input received from the user, the IVA generates responses by displaying pre-defined animations and utterances. Next, in order to answer the central research question, a between-subject experiment was conducted, to investigate the effects of the IVA in a consequential vs. (non-consequential) control condition on users' anxiety and (physiological and self-reported) stress levels. In the consequential condition, the IVA's behavior was physically threatening, as participants were told that it could activate a shock-device around the participant's finger. However, it is important to note that there was no actual shock. In the control condition, no consequences were involved in the interaction with the IVA. Effects on participants' levels of anxiety and stress were measured, and the role of presence and perceived believability of the IVA was assessed.

The remainder of this paper is structured as follows. In Section 2, the recent literature on believable agents is briefly discussed, with an emphasis on the role of virtual bad guys. Next, in Section 3 we put forward a theoretical framework about the expected effect of consequential virtual agents on user experience, resulting in a number of research hypotheses. In Section 4, we present the design of the experiment we conducted to test these hypotheses, and the results are provided in Section 5. The paper concludes in Section 6 with a discussion of the implications of these results.

2 BELIEVABLE VIRTUAL BAD GUYS

The Media Equation theory, proposed in 1996 by Reeves and Nass, states that people have an innate tendency to treat computers (and other media) as if they were real humans [34]. This tendency has profound consequences for the way people interact with computers, and in particular with IVAs. For instance, people often automatically ascribe personality characteristics and emotional states to these agents, which shapes the way they interact with them.

Nevertheless, this does not mean that every application based on human-agent interaction is guaranteed to have the desired effect. The literature on user experience with IVAs is extensive and at some points ambiguous, and the extent to which IVAs are really perceived as human-like seems to depend on many characteristics, including graphical and kinematic factors, but also behavioral and cognitive aspects of the agent [15, 26].

An important concept in this regard is *believability*, a property of virtual agents that refers to the extent to which they provide the illusion of being alive [3]. If the IVA represents a human character, believability refers to extent that users attribute unique human characteristics (addressed as "human essence" [25], i.e., intelligence, intentionality, emotions, etc.) to the agent. The more users find IVAs believable, however, the less they are inclined to think of them as artificially constructed beings. In [10], believability is defined by three dimensions, namely aesthetic, functional, and social qualities of agents, which can be related, respectively, to the agent's body (its physical appearance), mind (the mechanisms that drive its behavior), and personality (the traits that determine its interaction style). Particularly creating a believable personality is not easy. This aspect is typically assumed to be related to socio-emotional properties such as personality, attitudes and affect. The challenge to incorporate such properties within computational systems resulted in the mid-1990s in the emergence of the affective computing field [33]. Indeed, since that time, there has been a significant expansion in research on computational models of emotion (see [28] for an overview), resulting into the development of increasingly believable IVAs.

Nevertheless, only a small fraction of this research focuses explicitly on IVAs with negative emotions. In most cases, generic computational models of emotion are used. For instance, one of the most influential approaches is EMA [27], a computational model that formalizes the main assumptions behind appraisal theory [24]. Although such models could be used to generate negative emotional states like 'anger' at appropriate moments, they do not focus on the resulting behavior that is required to make users actually feel threatened by the agent.

Other research has focused more explicitly on the impact of emotional agents on humans in interpersonal settings. For example, the Sensitive Artificial Listener paradigm enables researchers to investigate the effect of agents with different personalities on human interlocutors. Studies using this paradigm have provided evidence that IVAs with an angry attitude indeed trigger different (subjective and behavioral) responses than agents with other personalities [39]. Along the same lines, de Melo and colleagues found that IVAs expressing anger (in terms of utterances and facial expressions) lead human negotiation partners to make larger concessions [11]. More recently, Blankendaal et al. demonstrated that an IVA showing aggressive behavior towards human interlocutors (in terms of shouting and insulting the human) in a 2D environment was able to trigger a significant physiological stress response, measured in terms of increased skin conductance levels [6].

Based on these studies it can be concluded that virtual bad guys can trigger certain subjective, behavioral, and even physiological responses in human interaction partners. However, whether these responses are comparable with the stress response that people experience during a real-life encounter with an aggressive individual is debatable. For instance, although the aggressive agent developed in [6] triggered a physiological response, this response was found to be significantly lower than a response triggered by an aggressive human. Another recent study, in which the authors attempted to make virtual bad guys more believable by incorporating haptic feedback (realized through a vibrating vest) in a virtual reality environment, led to the tentative conclusion that the stimuli used were of insufficient strength to make the IVA truly believable [16].

Some more encouraging results are reported in [36], in which the believability of an aggressive avatar in virtual reality is enhanced by using haptic feedback in the form of a lightweight exoskeleton. In a scenario in which the avatar touched and punched an avatar of the user, the haptic feedback was found to trigger significantly higher subjective and physiological (skin conductance) responses. However, the relatively low number of participants (16) and the within-subjects design of this study make it hard to generalize these results. Additionally, as this scenario did not involve any communication, the implications for interaction with truly 'social' agents that talk to (and have the ability to verbally threaten) the user remain to be explored.

To conclude, it is not easy to create a setting involving virtual bad guys that have the ability to physically threaten their human interlocutors in such a way that they are actually perceived as threatening in a human-like manner.

3 THEORETICAL FRAMEWORK

The problem of designing believable and stressful virtual bad guys relates to the general problem of mediated environments that people, during usage, might stay cognitively aware of their mediated nature and, thus, maintain a psychological distance to the depicted events [9]. This mechanism also applies to highly immersive VR environments. Recent dual-processing models of mediated reality [18, 23] suggest that although immersive VR environments might automatically "feel real" to users, they nevertheless might be simultaneously appraised as something abstract or artificial, as users continue to "know that this is not real". However, as it is well known from research, e.g., on horror movies, users' "knowing that this is not real" provides a powerful cognitive tool to dismantle or reappraise intense negative affect during exposure [38]. If users stay aware they encounter a benign threat [31] and actually are in a protected situation, unpleasant experiences that are automatically triggered during exposure, like intense distress or anxiety, might be quickly reversed into pleasurable thrill or excitement [1]. Accordingly, users might even appraise an encounter with an aggressively acting IVA –a virtual bad guy– in a highly realistic virtual environment as something playful or even enjoyable, rather than distressful and anxiety-evoking. This, in turn, would severely constrain the effectiveness of stressful training simulations.

The present approach pursues the idea that in order to induce stronger stress and anxiety responses in users that encounter a bad IVA, the protective layer that results from their cognitive awareness that "this is not a real threat" needs to be weakened. More specifically, we follow the idea that users would be less certain of being in a benign and protected situation if they had reason to believe that the aggressive IVA had the capacity to physically harm them by providing (mild) electric shocks, i.e., if they face a consequential IVA. Stress can be defined as "a physiological reaction of the autonomic nervous system to a threatening stimulus" [8, p. 302]. It is characterized by high but unpleasant arousal levels [37]. Anxiety, too, is a response to perceived threat or danger and can be "considered similar and perhaps identical to the reaction of fear" [30, p. 234]. Stress and anxiety are triggered by the perception of an immediate physical threat. Accordingly, we hypothesized that interacting with an ostensibly physically threatening IVA induces stronger states of stress and anxiety in users than interacting with a non-physically threatening virtual character (H1).

The postulated effect of an ostensibly physically threatening IVA on stress and anxiety might build on (i.e.., might be mediated by) both users' sense of spatial [42] and social presence [5] as well as their perceived believability of the character [2]. Spatial presence refers to users' subjective experience of being physically located in a virtual environment [42]. Social presence implies that users feel a sense of co-presence, mutual awareness and attention with the virtual other [5]. If users feel spatially and socially present in the virtual environment, psychological distance is reduced and unfolding events and encountered characters should be perceived as being of greater significance for their immediate wellbeing [18]. Accordingly, the more intense users' presence experience, the stronger should be their stress and anxiety levels if encountering an aggressive IVA. At the same time, if users believe that the encountered IVA might actually physically harm them, they might pay more attention to the encounter, become more cognitively involved, and, as a consequence, feel more present in the virtual environment [42] than users who have no reason to believe in such a threat. Accordingly, we hypothesized that the effect postulated in H1 might be mediated by, i.e., be partly due to a more intense experience of spatial and social presence among users encountering a consequential IVA that assumingly could physically harm them (H2).

Furthermore, we reasoned that the effect postulated by H1 builds on perceived believability. In the present approach, we understood believability as a perception of users that the IVA is alive and human-like, i.e., possesses "human essence" [25]. This perception, also addressed in the literature as anthropomorphism [13] or "seeing human", entails that users automatically ascribe human-like features to the IVA, such as intelligence and a mind of his/her own, agency and intention, and inner sentiment or feelings. We assumed that the more users find the aggressive IVA believable, the more they should feel anxious and stressed. With regards to the present experiment, users (in the consequential condition) should believe that the encountered IVA is able to actually physically harm them. Therefore, they might ascribe greater agency and intentionality, and, overall, greater human essence, to the IVA as compared to users in the non-consequential (control) condition. Accordingly, we hypothesized that the effect postulated in H1 might be mediated by, i.e., be partly due to a greater believability of the IVA in the consequential as compared to the non-consequential (control) condition (H3).

A graphical overview of the three hypotheses used in this study is shown in Figure 1. In this figure, the different (independent, mediating and dependent) variables used are depicted as rounded rectangles, and the hypothesized dependencies as arrows.

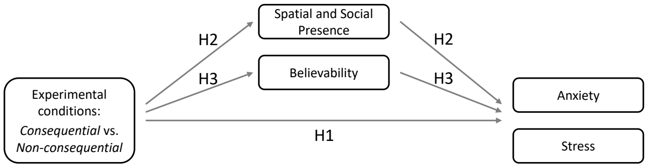
Spatial and Social Presence
Believability
Experimental conditions: *Consequential* vs. *Non-consequential*
H2
H3
H2
H3
H1
Anxiety
Stress

Figure 1: Hypotheses used in the study.

4 METHOD

This section describes the experiment that was conducted to test the hypotheses formulated in the previous section.

4.1 Participants

A convenience sample of 52 healthy adults (33 male, 19 female) was recruited, most of which were academic students. They were recruited through word-of-mouth and social media in the personal networks of the researchers. In addition, flyers were handed out in several buildings of VU University. Participants could enroll themselves for the experiment via an online subscription platform. Their average age was 25.06 years (*SD* = 7.57). The study protocol was approved by the ethics committee of the Faculty of Social Sciences. Participants had to pass a health check[1], and gave written informed consent for their participation in the study. They received a gift voucher of 5 Euros for their participation.

4.2 Experimental Design

In order to test our hypotheses, we conducted a one-factorial between-subjects experiment in which we compared the effects of a consequential version of an aggressive IVA (experimental condition) to a non-consequential version of the same IVA (control condition) on participants' self-reported anxiety and stress levels, as well as their physiological stress response. Participants were randomly assigned to either one of the two experimental conditions. Condition one, the consequential condition (N=26), included a (presumed) physical threat, where the participants were told that the IVA was able to shock them through a device that was attached to their finger (which actually is a skin conductance sensor, see Section 4.5). Condition two was the control condition (N=26), without any mention of the physical threat. Participants in this condition also wore the skin conductance sensor, but they were informed of the real purpose of this device. Both conditions involved the same interactive scenario between the participant and the IVA (see Section 4.3), in which the IVA's behavior was aggressive. Hence, the only difference was the fact that participants in the consequential condition believed that the IVA could give them a small shock if its levels of aggression would exceed a certain limit.

To ensure that participants in the consequential condition believed that they could actually get a small shock, they were shown a fake video of another person receiving such a shock. This video showed the exact setup of the experiment, but was fake in the sense that the person in the video acted as if she felt a shock although she did not actually receive it. In addition, participants in this condition were told that experiencing the shock was comparable to touching an electric cattle fence[2].

4.3 Tasks

Participants were asked to engage in a virtual reality scenario (displayed on a Head Mounted Display) that took place inside a pub. While playing the scenario, they were sitting at a table, making it impossible for them to walk around. However, they could move their head and look around in the virtual pub. The main reason for having the participants seated in the VR environment was to make their experience as realistic as possible, by avoiding any inconsistencies between their actual posture and their perspective in the virtual environment. Before starting the experiment, participants were told that they had an appointment with a friend to have a drink together, and they were instructed to look for that friend in the virtual environment. Within the virtual pub, a number of avatars were present. One of them was our 'virtual bad guy'. This avatar was initially standing at the bar, but as soon as our participant looked at him (which could be calculated based on the camera position using ray casting techniques), he would approach the participant and start a conversation. An illustration of a conversation with the virtual bad guy is shown in Figure 2.

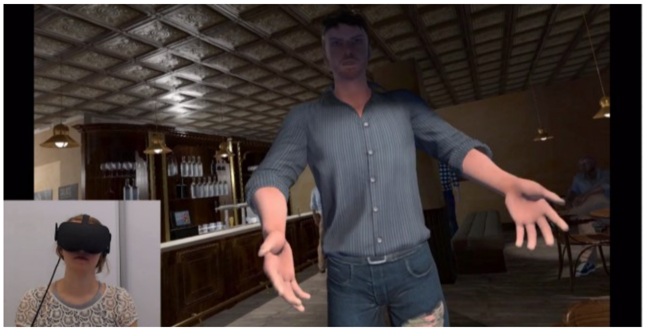

Figure 2: Virtual bad guy interacting with a participant.

[1] Two participants dropped out (not included in the N=52).

[2] As part of the post-experiment survey, we checked whether participants in this condition indeed believed that they were wearing was a shock device.

The approach used to enable a conversation between the agent and the participant was inspired by the Sensitive Artificial Listener paradigm [39], and resulted in the following scenario. The IVA started the conversation by saying 'Hey, what are you looking at?' with an angry voice. After that, the conversation continued in a turn-based manner, where the participant could interact with the IVA using free, open-ended speech. As soon as the participant started talking, the system detected when (s)he had finished a sentence, which was the trigger for the IVA to introduce the next sentence. All sentences (and corresponding animations) produced by the IVA were fixed, and developed in advance.

The entire scenario consisted of six interactions (i.e., 6 sentences spoken by the IVA and 6 responses by the participant). During these interactions, the verbal and non-verbal behavior of the IVA gradually increased in terms of aggression level, no matter how the participant reacted. Although the agent followed a fixed script, the sentences were constructed in such a way that the agent always seemed to respond to what the participant said (see Appendix A for the complete script). Halfway the scenario, the agent bent over the table at which the participant was sitting, taking an even more threatening posture. During the last sentence, the agent moved its arm as if it would hit the participant.

4.4 Measures

Anxiety. A pre-post design was applied to assess the change in participants' anxiety and stress levels before vs. after encountering the IVA. Participants' level of anxiety was assessed based on a translation of a self-report measure by Marteau and Bekker [29] that we applied before and after the virtual encounter. The original scale assessed state anxiety based on six items (e.g., "I'm worried") on a 4-point scale reaching from 1 (not at all) to 4 (very much). We added another item directly assessing anxiety ("I feel anxious") to the scale. After dropping one item ("I feel angry"), the measure yielded high internal reliability ($\alpha_{t1} = .79$, $\alpha_{t2} = .84$).

Stress. We assessed participants' stress response based on self-reports and physiological data. Self-reported stress was assessed based on the single-item affect grid measure [37]. Participants reported on their current affective state selecting a matching level of arousal (ranging from sleepy to aroused) and valence (ranging from negative to positive) on a 9x9 grid. Stress was indicated by higher arousal and more negative valence scores. We assessed self-reported stress with the affect grid before and after the virtual encounter.

In addition, physiological stress was assessed based on participants' skin conductance levels, also called electrodermal activity (EDA) [14]. We assessed EDA at baseline prior to the VR exposure and continuously during the virtual encounter. In the present paper, we apply a pre-post logic by comparing participants' EDA response at baseline vs. at the end of the VR experience, in which the interaction with the aggressive IVA reached its climax.

Presence. We assessed self-reported spatial and social presence as potential mediators. Both concepts were applied as a retrospective measure after the VR experience. Spatial presence was assessed based on the 4-item Spatial Presence Experience Scale (SPES [20], example item "I felt as though I was physically present in the environment of the presentation"). Participants rated items on a 5-point answering scale ranging from 1 ("I do not agree at all") to 5 ("I totally agree"). After dropping one item, the scale yielded a high internal reliability, $\alpha = .83$.

Social presence was assessed based on the Experience of Parasocial Interaction Scale (EPSI [19]), a 6-item measure that reflects participants' subjective sense of mutual awareness, mutual attention, and mutual adjustment with the IVA. These perceptions can be considered the core of the social presence experience [5]. Participants answered items on a 7-point scale ranging from 1 ("I do not agree at all") to 7 ("I totally agree"). The scale yielded high internal reliability, $\alpha = .87$.

Believability. We assessed believability as another potential mediator based on a 5-item short scale of "perceived human essence" developed by Hartmann [19]. The scale was applied as a retrospective measure after the VR experience. Participants rated items expressing the subjective perception that the virtual character seemed to be alive, have his own feelings, a personality, his own life, and a soul on a 5-point scale ranging from 1 ("not at all") to 5 ("very much"). The scale proved to be internally reliable, $\alpha = .85$.

4.5 Material

The experiment took place in a quiet room in which only the participant and the experimenter were present. The experimenter kept a few meters distance from the participant, to give the participant sufficient privacy to play the scenario, while still being available in case (s)he needed help. The room also contained a desk with the computer that hosted the virtual environment and a chair for the participants to sit on (see Figure 3).

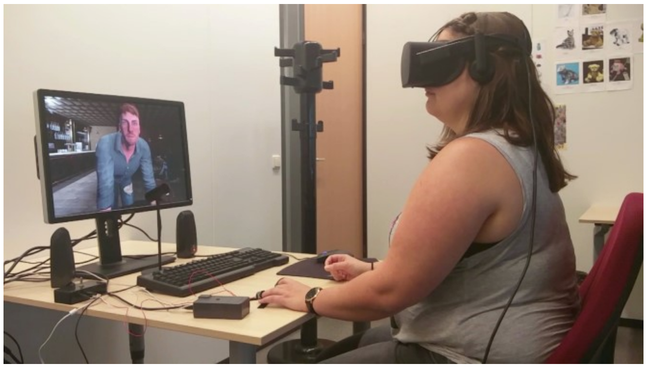

Figure 3: Setup of the experiment.

The virtual environment was presented to the user by means of a Head Mounted Display, in this case the Oculus Rift[3]. Using an advanced high-quality virtual environment and a Head Mounted Display required a high-end gaming computer with a strong graphics card to ensure smooth performance for an optimally effective virtual environment. The computer used an Intel i7-4630 CPU with 16GB DDR4 memory, a 500GB SSD and a Nvidia GTX-

[3] https://www.oculus.com/rift/

1080 graphics card with 8GB of memory. Sound was provided through the headphones of the Head Mounted Display. In addition, participants wore a custom-made skin conductance sensor. This sensor used two simple electrodes that were strapped to two non-adjacent fingers of the left hand. A Phidget I/O board was used to record the conductance and communicate with the virtual environment. Code in the virtual environment calculated the skin conductance in microSiemens (μS) values.

4.6 Virtual Environment

The virtual environment was developed in Unity 3D (version 5)[4]. A ready-made model from the Unity Asset Store was purchased for the pub environment used in the experiment. This model was further adapted in order to suit the needs of this research. Atmosphere was added by including special lighting and additional properties on the virtual stage. All the humanoid agents in the virtual environment were generated using the iClone Pipeline software (version 6)[5]. The Character Creator[6] was used to generate realistic and unique human agents. iClone itself was used to create the body animations and lip-sync movements.

3DXchange[7] was used to convert the agents including their animations into FBX format that could be imported into Unity. Within Unity, the non-interactive characters were scripted using C#, looping animations and speech to create a livelier atmosphere in the pub. The interactive "virtual bad guy" was scripted separately for more advanced actions. This agent had a larger set of animations and speech, plus the ability to time its reactions based on the speech of the participants. More specifically, the agent could monitor if the participant was speaking. If the participant did speak, the agent would wait until the participant stopped, allowing for small pauses in speech (of 1 second), or until a maximum amount of time (of 10 seconds) had elapsed.

4.7 Procedure

After entering the room, participants were asked to fill out the first part of the survey. This survey was used to measure participants' pre-exposure states of anxiety and stress. In addition, it included a general cover story ("we are interested in your personal experience with a virtual agent") and an informed consent form. All participants were explicitly asked for health problems; for instance, participants with a heart condition were not allowed to participate in the experiment.

At the end of the first part of the survey and during the virtual reality experience, electrodermal activity was measured. The baseline EDA response was assessed while participants watched a short video including a peaceful aquarium with relaxing background music. Afterwards, participants were briefed about the follow-up VR session. In the consequential condition, the briefing included the presentation of the fake video about the electric shock, as described in Section 4.2. Participants then put on the virtual reality equipment and started the scenario. After the scenario, participants could take off the head mounted display, headphones and electro dermal activity device and fill out the second part of the survey. This survey assessed post-exposure states of anxiety and stress[8]. Finally, participants were debriefed and received a gift voucher as a reward for their participation.

5 RESULTS

We analyzed the data with the IBM SPSS Statistics software package. H1 was tested in four separate mixed ANOVAs, in which either pre-post self-reported anxiety, arousal or valence, or physiological stress were entered as a repeated measures factor, and the experimental factor (consequential vs. non-consequential control) was entered as between-subjects factor. The three ANOVAS examining self-report data yielded a significant increase in anxiety, $F(1,50) = 681.61$, $p < .01$, $\eta_p^{2 =}$ .93, arousal, $F(1,50) = 60.70$, $p < .01$, $\eta_p^{2 =}$ .55, and decrease in valence, $F(1,50) = 56.35$, $p < .01$, $\eta_p^{2 =}$ .53, in the pre-post comparison. However, these effects did not differ depending on whether participants were in the experimental or control condition, $F_{anxietyXcondition}(1,50) = 0.86$, *n.s.*, $F_{arousalXcondition}(1,50) = 0.17$, *n.s.*, $F_{valenceXcondition}(1,50) = 0.04$, *n.s.* In contrast, the ANOVA on physiological stress revealed a significant increase of stress over time, $F(1,50) = 101.85$, $p < .01$, $\eta_p^{2 =}$ .57, that was qualified by the experimental factor, $F_{phys.stressXcondition}(1,50) = 5.05$, $p < .05$, $\eta_p^{2} =$ .09. Hence, physiological stress increased significantly more strongly in the consequential condition as compared to the non-consequential condition (see Figure 4). In summary, H1 was supported only for physiological stress. In addition, results show that the aggressive IVA resulted in greater self-report anxiety and stress among users independent of the experimental manipulation.

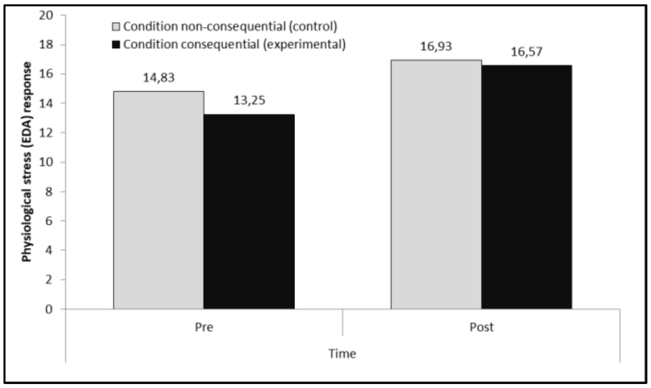

Figure 4: Increase of physiological stress (EDA) response in microSiemens, depending on condition (N = 52).

H2 and H3 were tested in eight separate mediation regression analyses that utilized the PROCESS macro by Hayes [21]. Each hypothesis was tested four times, by employing the post-measure of self-reported anxiety, arousal, valence, or physiological arousal

[4] https://unity3d.com/

[5] http://www.reallusion.com/iclone/

[6] http://www.reallusion.com/iclone/character-creator/

[7] http://www.reallusion.com/iclone/3DXchange.html

[8] Note that participants were asked to report the anxiety and stress that they experienced *during* the interaction with the IVA, not after the interaction.

as the dependent variable. In all analyses, the experimental factor represented the independent variable. The test of H2 included spatial and social presence as two parallel mediators, the test of H3 included believability as a mediator. In all tests, we controlled for the effect of the respective pre-exposure state on the dependent variable.

Contrary to H2 and H3, none of the eight tests yielded a significant indirect effect (estimated based on 1000 bootstrap samples) that would indicate a mediation effect. As a general pattern of results, we observed that whereas the mediators were significantly related to the outcomes, they did not significantly differ between experimental conditions. Accordingly, contrary to our expectations, participants in the consequential condition did not feel a significantly stronger sense of spatial and social presence, or found the IVA more believable, as compared to participants in the non-consequential control condition. However, as follow-up zero-order correlational analyses confirmed, feelings of presence and believability were significantly associated with self-reported post-exposure anxiety and the valence of participants' affect, and, albeit to a lesser degree, their self-reported arousal levels (see Table 1).

Table 1: Correlations (Pearson's r) between spatial and social presence, believability vs. post-measures.

	Self-report data			Physiological data
	Anxiety	Arousal	Valence	Arousal (EDA)
Spatial Presence	.29*	.23^{+}	-.25^{+}	-.04
Social Presence	.45*	.06	-.37*	-.06
Believability	.44*	.24^{+}	-.33*	.03

Note. ^{+}p <.10, $^{*}p$ <.05, $^{**}p$<.01

Accordingly, independent of the experimental manipulation, those who found the IVA more believable, and experienced greater spatial and social presence during the VR session, also reported feeling more anxious and more stressed afterwards.

6 DISCUSSION AND CONCLUSION

Whilst the body of literature on intelligent virtual agents is growing rapidly, the vast majority of the existing papers concentrates on agents with a positive stance towards the user. Instead, the current paper explored the concept of virtual bad guys that have the ability to physically threaten human beings. Based on the assumption that virtual bad guys can only be taken seriously if they have an effect in the real world, we investigated the impact of IVAs which people believe could physically harm them.

The results of the study indicate, first of all, that both the consequential and the non-consequential IVA were successful in increasing the participants' physiological and self-reported stress and anxiety. This is an encouraging finding, because stress is claimed to be an important requirement for various effective applications (e.g., for resilience training [32]). Moreover, when looking at the differences between the two conditions, the consequential agent was found to increase physiological stress more than the non-consequential agent, which partly confirms H1. However, this finding did not hold for the self-reported stress and anxiety. A possible explanation of this discrepancy could be that, since the participants experienced only one of the two conditions, they had no 'anchor' when answering the self-report questions. For instance, when rating a statement like "I feel anxious", people may have different interpretations of the value "very much". Another explanation could be that, since people had to report their stress and anxiety after the virtual encounter, they already partly forgot how they felt.

The analysis of H2 indicated that, contrary to our expectations, participants in the consequential condition did not feel a stronger sense of (spatial and social) presence than participants in the non-consequential condition. Similarly, they also did not find the IVA more believable, which was postulated in H3. However, an interesting additional finding is that higher experienced presence and reported believability of the IVA were correlated with higher self-reported anxiety and stress.

To conclude, the current study presented some initial evidence that it is possible to trigger (at least physiological) stress using consequential virtual agents, which may be useful for a variety of applications. However, more research is definitely needed to gain a better understanding of the different factors that play a role in this process. For example, the main variable that was manipulated in the current experiment was users' *expectation* of threat, but not actual threat (after all, no actual shocks were administered). It would therefore be interesting to investigate the difference in impact between IVAs that are believed to harm people and IVAs that can actually evoke sensations associated with physical harm, e.g., by using haptic technology as in [16,36].

Another factor that is worth exploring further is the coupling between the threatening device in the real world (in our case: the presumed shock device) and the threat in the virtual world (i.e., the agent moving its arm to hit the participant). Some of the null results may be explained by the fact that these two concepts were not very closely connected in the present study. For instance, the agent did not carry a virtual shock device, nor did the participants see a replica of their hands in the VR environment. Such elements have been shown to contribute positively to user experience, for instance in a VR variant of the 'rubber hand illusion' experiment, which indicated that a virtual hand can be made to feel part of one's own body [40]. In follow-up experiments, these elements could be used to make participants more aware of the threat.

Besides physical threat, it is also worthwhile to investigate if there are any other mechanisms via which agents could threaten human beings, and what would be the consequences of such types of threat. One may think for instance about a more "psychological" type of threatening, e.g., in the form of agents that carry sensitive personal information about the user, and threaten to disclose this. Similarly, in the context of a game where players can win or lose money, agents that have the power to decide about the payoff may use this as part of their threats.

A last direction for future research would be a more detailed investigation of potential moderating variables such as personality or other individual characteristics. As an example, people who are more resistant to stress in general may be less susceptible to the manipulations explored in this research. Also, a potentially moderating factor is the predictability of the IVA. The agent used in the current study was rather predictable in the sense that it followed a fixed script of pre-recorded sentences, which may have resulted in a less realistic experience. Instead, it would be interesting to investigate the effect of IVA's that are more flexible in generating their behavior. To this end, various Artificial Intelligence techniques could be used, varying from natural language analysis to social signal processes techniques. In our current work in progress, we are already experimenting with this by developing an IVA that adapts its aggression level during the interaction to the EDA measurements of the user.

Finally, we would like to emphasize that, however fascinating, the concept of consequential virtual agents brings along some complicated ethical issues, which are even more important given recent discussions about the 'risks of AI' [17]. For the current project, an exploration of the ethical boundaries and implications was an explicit part of the research activities, and we think that any potentially harmful effect was minimized by the way the experiment was designed. However, the current study should by no means be seen as a plea for technology that do actual harm to human beings. On the contrary, we hope that it can serve as an example of a project that studies an ethically complicated topic like aggression in a controlled and responsible manner.

ACKNOWLEDGMENTS

This project was supported by http://www.networkinstitute.org/.

REFERENCES

[1] Andrade, E. B. and Cohen, J. B. (2007). On the consumption of negative feelings. Journal of Consumer Research.

[2] Bartneck, C., Kulić, D., Croft, E., and Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. International journal of social robotics, 1(1):71-81.

[3] Bates, J. (1994). The role of emotions in believable agents. Communications of the ACM, 37(7):122-125.

[4] Beskow, J., Peters, C., Castellano, G., O'Sullivan, C., Leite, I., and Kopp, S. (2017). Intelligent Virtual Agents. 17th International Conference. Proceedings. Springer LNCS, vol. 10498.

[5] Biocca, F. and Harms, C. (2002). What is social presence? In F. Gouveia and F. Biocca (Eds.), Proceedings of Presence 2002. Porto, Portugal: University of Fernando Pessoa Press.

[6] Blankendaal, R., Bosse, T. Gerritsen, C., de Jong, T., and de Man, J. (2015). Are aggressive agents as scary as aggressive humans? In Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2015, Istanbul, Turkey, May 4-8, 2015, pages 553-561.

[7] Bosse, T., Gerritsen, C., and de Man, J. (2016). An intelligent system for aggression de-escalation training. In: Proceedings of the 22nd European Conference on Artificial Intelligence, ECAI 2016. IOS Press.

[8] Bosse, T., Gerritsen, C., de Man, J. and Stam, M. (2014). Inducing anxiety through video material. Communications in Computer and Information Science. HCI International 2014 - Posters' Extended Abstracts:301-306.

[9] Cupchick, G.C. (2002). The evolution of psychical distance as an aesthetic concept. Culture Psychology, 8:155-187.

[10] De Angeli, A., Lynch, P., and Johnson, G. (2001). Personifying the e-market: A framework for social agents. In: M. Hirose (Ed.), Proceedings of Interact 2001. IOS Press, pp. 198-205.

[11] De Melo, C.M., Carnevale, P.J., and Gratch, J. (2011). The effect of expression of anger and happiness in computer agents on negotiations with humans. In 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2011), Taipei, Taiwan, May 2-6, 2011, Volume 1-3, pages 937-944.

[12] DeVault, D. et al. (2014). Simsensei kiosk: a virtual human interviewer for healthcare decision support. In: Int. conference on Autonomous Agents and Multi-Agent Systems, AAMAS '14, Paris, France, May 5-9, pp. 1061-1068.

[13] Epley, N, Akalis, S, Waytz, A, and Cacioppo, J.T. (2008). Creating social connection through inferential reproduction: Loneliness and perceived agency in gadgets, Gods, and greyhounds. Psychological Science, 19:114-120.

[14] Farrow, T.F.D., Johnson, N.K., Hunter, M.D., Barker, A.T., Wilkinson, I.D., and Woodruff, P.W.R. (2013). Neural correlates of the behavioral-autonomic interaction response to potentially threatening stimuli. Frontiers in Human Neuroscience, 6.

[15] Funge, J., Tu, X., and Terzopoulos, D. (1999). Cognitive modeling: knowledge, reasoning, and planning for intelligent characters. Proc. of SIGGRAPH'99: 26th conference on computer graphics, ACM Press, New York, pp. 29-38.

[16] Goedschalk, L., Bosse, T., and Otte, M. (2017). Get Your Virtual Hands Off Me! - Developing Threatening Agents Using Haptic Feedback. In: Proceedings of the 29th Benelux Conference on Artificial Intelligence, BNAIC'17.

[17] Gurkaynak, G., Yilmaz, I., and Haksever, G. (2016). Stifling artificial intelligence: Human perils. Computer Law & Security Review 32, pp 749–758.

[18] Hartmann, T. (2011). Players' experiential and rational processing of virtual violence. In S. Malliet and K. Poels (Eds.), Vice City Virtue. Moral Issues in Digital Game Play (pp. 135-150). Leuven: Acco.

[19] Hartmann, T., and Goldhoorn, C. (2011). Horton and Wohl revisited: Exploring viewers' experience of parasocial interaction. Journal of Communication, 61: 1104-1121.

[20] Hartmann, T., Wirth, W., Schramm, H., Klimmt, C., Vorderer, P., Gysbers, A., Böcking, S., Ravaja, N., Laarni, J., Saari, T., Gouveia, F., and Sacau, A. (2016). The spatial presence experience scale (SPES): A short self-report measure for diverse media settings. Journal of Media Psychology, 28(1): 1-15.

[21] Hayes, A. F. (2012). PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling [White paper]. Retrieved from http://www.afhayes.com/public/process2012.pdf.

[22] Ieronutti, L and Chittaro, L. (2007). Employing virtual humans for education and training in X3D/VRML worlds. Computers & Education, 49(1), 93–109.

[23] Konijn, E.A. and ten Holt, J. M. (2011). From noise to nucleus. Emotion as a key construct in processing media messages. In K. Döveling, C. van Scheve, and E.A. Konijn (Eds.), The Routledge Handbook of Emotions and Mass Media (pp. 37 - 59). New York: Routledge.

[24] Lazarus, R.S. (1991). Cognition and motivation in emotion. American Psychologist, 46:352-367.

[25] Leyens, J. P., Rodriguez, A. P., Rodriguez, R. T., Gaunt, R., Paladino, P. M., Vaes, J., and Demoulin, S. (2001). Psychological essentialism and the attribution of uniquely human emotions to ingroups and outgroups. European Journal of Social Psychology, 31:395–411.

[26] Magnenat-Thalmann, N. and Thalmann, D. (2005). Virtual humans: thirty years of research, what next? The Visual Computer 21, 12, 997-1015.

[27] Marsella, S. and Gratch, J. (2009). EMA: A process model of appraisal dynamics. Cognitive Systems Research, 10(1):70-90.

[28] Marsella, S., Gratch, J., and Petta, P. (2010). A blueprint for an affectively competent agent: Crossfertilization between Emotion Psychology, Affective Neuroscience, and Affective Computing., chapter Computational Models of Emotion. Oxford University Press.

[29] Marteau, T. M. and Bekker, H. (1992). The development of a six-item short-form of the state scale of the Spielberger State—Trait Anxiety Inventory (STAI). British Journal of Clinical Psychology, 31(3): 301-306.

[30] Martin, B. (1961). The assessment of anxiety by physiological behavioral measures. *Psychological Bulletin, 58*(3):234-255.

[31] McGraw, A.P. and Warren, C. (2014). Benign violation theory. Encyclopedia of Humor Studies:75-77.

[32] Nieuwenhuys, A. and Oudejans, R.R.D. (2011). Training with anxiety: short- and long-term effects on police offcers' shooting behavior under pressure. Cognitive Processing, 12(3):277-288.

[33] Picard, R. (1997). Affective Computing. MIT Press.

[34] Reeves, B. and Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people and places. Stanford, CA: CSLI Publications.

[35] Rizzo, A.A., Reger, G., Gahm, G., Difede, J., and Rothbaum, B.O. (2008). Virtual Reality Exposure Therapy for Combat Related PTSD. In: Shiromani, P., Keane, T., and LeDoux, J. (eds.), Post-Traumatic Stress Disorder: Basic Science and Clinical Practice, Springer Verlag.

[36] Ruffaldi, E., Barsotti, M., Leonardis, D., Bassani, G., Frisoli, A., and Bergamasco, M. (2014). Evaluating virtual embodiment with the alex exoskeleton. In: 'Haptics: Neuroscience Devices Modeling and Applications', Springer Verlag, pp. 133-140.

[37] Russell, J. A., Weiss, A., and Mendelsohn, G. A. (1989). Affect grid: A single-item scale of pleasure and arousal. Journal of Personality and Social Psychology, 57(3): 493-502.

[38] Schramm , H. and Wirth, W. (2008). A case for an integrative view on affect regulation through media usage. Communications: The European Journal of Communication Research, 33:27–46.

[39] Schroeder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., ter Maat, M., McKeown, G., Pammi, S., Pantic, M., Pelachaud, C., Schuller, B., de Sevin, E., Valstar, M., and Woellmer, M. (2012). Building autonomous sensitive artificial listeners. IEEE Transactions on Affective Computing, 3(2):165-183.

[40] Slater, M., Perez-Marcos, D., Ehrsson, H.H., and Sanchez-Vives, M.V. (2008). Frontiers in Human Neuroscience 2(6).

[41] Swartout, W., Gratch, J., Hill, R.W., Hovy, E., Marsella S., Rickel, J., and Traum, D. (2006). Toward Virtual Humans. AI Magazine 27(2):96-108.

[42] Wirth, W., Hartmann, T., Boecking, S., Vorderer, P., Klimmt, P., Schramm, H., Saari, T., Laarni, J., Ravaja, N., Gouveia, F. R., Biocca, F., Gouveia, L. B., Rebeiro, N., Sacau, A., Jäncke, L., Baumgartner T., and Jäncke, P. (2007). A Process Model of the Formation of Spatial Presence Experiences. Media Psychology, 9:493-525.

[43] Zoll, C., Enz, S., Schaub, H., Aylett, R., and Paiva, A. (2006). Fighting Bullying with the Help of Autonomous Agents in a Virtual School Environment. In: Proc. of the 7th International Conference on Cognitive Modelling (ICCM).

APPENDIX A

Script of the scenario (translated from Dutch).

agent	'Hey, what are you looking at?'
user	…
agent	'Am I wearing your clothes, or what?'
user	…
agent	'Hey, are you listening to what I'm saying?'
user	…
agent	'You just want me to get angry, huh?'
user	…
agent	'One more remark and I'm gonna hurt you!'
user	…
agent	'OK, that's enough!!!'