Beyond Believability: Quantifying the Differences between Real and Virtual Humans

Celso M. de Melo[1], and Jonathan Gratch[2]

[1]USC Marshall School of Business, Los Angeles, CA 90089-0808
demelo@usc.edu
[2]Institute for Creative Technologies, University of Southern California,
12015 Waterfront Drive, Building #4, Playa Vista, CA 90094-2536
gratch@ict.usc.edu

Abstract. "Believable" agents are supposed to "suspend the audience's disbelief" and provide the "illusion of life". However, beyond such high-level definitions, which are prone to subjective interpretation, there is not much more to help researchers systematically create or assess whether their agents are believable. In this paper we propose a more pragmatic and useful benchmark than believability for designing virtual agents. This benchmark requires people, *in a specific social situation, to act with the virtual agent in the same manner as they would with a real human*. We propose that perceptions of mind in virtual agents, especially pertaining to *agency* – the ability to act and plan – and *experience* – the ability to sense and feel emotion – are critical for achieving this new benchmark. We also review current computational systems that fail, pass, and even surpass this benchmark and show how a theoretical framework based on perceptions of mind can shed light into these systems. We also discuss a few important cases where it is better if virtual humans do *not* pass the benchmark. We discuss implications for the design of virtual agents that can be as natural and efficient to interact with as real humans.

Keywords: Believability, Mind Perception, Emotion, Virtual vs. Real Humans

1 Introduction

The intelligent virtual agents community has always been fascinated with building "believable" agents. These agents are meant to provide the "illusion of life" and support the audience's "suspension of disbelief" [1]–[4]. The notion emerged from the arts and was a natural reaction to the focus, at the time, artificial intelligence researchers placed on simulating proper reasoning, problem solving, and logical-analytical skill. This fresh new perspective led researchers to, among others, develop agents that were driven by personality and expressed emotion.

Believability was, nevertheless, left mostly underspecified. As researchers attempted to determine the requirements for achieving believable agents, it became clear that believability was hard to measure with any precision or reliability. Some researchers did attempt to refine the notion of believability [5], [6] but, ultimately, the concept remained prone to subjective interpretation and, consequently, hard to study from a scientific point of view.

In this paper we propose a more pragmatic, clearly defined, and useful benchmark than believability. The benchmark is that: *in a specific social setting, people behave*

with the virtual agent in the same manner as they would with a real human. In social decision making, this benchmark is achieved when, for instance, people are as fair, generous, or cooperative with virtual as with real humans. In a learning task, the benchmark is achieved when people learn as much and as efficiently with virtual as with real humans. In a therapy session, the benchmark is achieved when, for instance, people self-disclose as much with the virtual as with the real doctor. The benchmark is, thus, really a point in a continuum, where there are virtual agents that fall below it (probably the majority) and others that actually surpass it. Finally, in contrast to believability which originally came from the arts, achieving our benchmark can be informed by rigorous communication and psychological theories. Section 2 overviews these theories. Section 3 reviews critical work that demonstrates how these theories and benchmark can guide the design of virtual agents in various domains. Section 4 will, then, present our conclusions.

2 Theoretical Framework

Clifford Nass and colleagues were among the first to advance a general theory of how humans interact with machines [7]–[9]. The theory's main tenet is that to the extent that machines display social cues (e.g., interactivity, verbal and nonverbal behavior, filling of typically human roles), people will treat them in a fundamentally social manner. The argument is that people "mindlessly" treat computers that exhibit social traits like other people as a way to conserve cognitive effort and maximize response efficiency [8]. These automatic cognitive heuristics lead people to use the easily accessible social rules from human-human interaction and apply them in human-machine settings. To support their theory, they replicated in a human-computer context various findings from the human-human interaction literature. For instance, they demonstrated that people were polite to computers [10], treated computers that were perceived to be teammates better [11], and even applied gender and race stereotypes to computers [12].

These initial findings were so promising that they actually proposed that it was possible to replicate any finding in human-human interaction with computers:

"Findings and experimental methods from the social sciences can be applied directly to human-media interaction. It is possible to take a psychology research paper about how people respond to other people, replace the word 'human' with the word 'computer' and get the same results." ([7], pg. 28)

Thus, a strict interpretation of the theory suggests that, in social settings, people will treat machines – virtual agents included – just like real humans and, thus, immediately meet our proposed benchmark.

Subsequent studies, however, showed that, even though people treat machines in a social manner, people still make important distinctions between humans and machines. For instance, these studies showed that, in certain social settings, people experienced higher social presence [13], [14], inhibition [15], learning [16], flow [17],

arousal [18], [19] and engagement [14] with humans than machines. These kind of findings led Blascovich and colleagues [20], [21] to propose that social influence would be greater in machines, the higher the perceived mindfulness[1]. According to this view, thus, the higher the attributions of mind people make, the more likely machines are to pass our benchmark.

Research shows that people are, in fact, quite adept at anthropomorphizing – i.e., attributing human-like qualities, including mental states – to non-human entities [22], [23]. Perceiving mind in (human or non-human) others matters because, when we see mind in others, we attribute more responsibility, moral rights, and respect to others [24]. On the other hand, when we deny mind to others, we dehumanize, and consequently, discriminate others [25].

Recent research, furthermore, suggests that we perceive mind in others according to two core dimensions [26]: *agency*, the ability to act and plan; and, *experience*, the ability to sense and feel emotion. When we deny agency to others [25], [27], we treat others like "animals" that possess primitive feelings, but no higher reasoning skills. When we deny experience to others, we treat others like "cold emotionless automata" ("business people"). Accordingly, in a survey involving thousands of participants, Gray, Gray, and Wegner [26] showed that: (a) adult humans were rated high in perceived agency and in perceived experience; (b) animals and babies rated high in experience, but low in agency; finally, (c) machines rated high in agency but very low in experience. According to this view, therefore, machines are unlikely to pass our benchmark, at least by default, because people perceive less mind in machines than humans, especially pertaining to perceptions of experience. This research, thus, goes one step further than Blascovich et al., in that it proposes a structure for perceiving mind in human and non-human others. The implication is that, the higher the perceived agency and experience in virtual humans, the more likely they are to pass our benchmark.

3 Empirical Evidence

In this section we present several studies that compare people's behavior with machines vs. humans. This review is not meant to be exhaustive, but representative of computational systems – many of which involving virtual agents – that failed, passed, and even surpassed our benchmark. We also present cases where it is actually better *not* to pass the benchmark, i.e., where the goal is to develop virtual agents that should not act like humans. We take particular care to frame all these systems within the theoretical framework mentioned in the previous section.

---

[1] Blascovich et al. (2002, 2013) used the term "agency" instead of "mindfulness"; however, this use of the term conflicts with its use in the mind perception literature. In this paper, we adopt the latter definition of agency.

3.1 Systems That Are Not as Good As Humans

Neuroeconomics is an emerging field that studies the biological basis of decision making in the brain [28]. To accomplish this, researchers compared people's behavior with humans vs. computers. This evidence reveals that people tend, by default, to reach different decisions and show different patterns of brain activation with machines in the exact same decision making tasks, for the exact same financial incentives, when compared to humans [29]–[36]. For instance, Gallagher et al. [29] showed that when people played the rock-paper-scissors game with a human there was activation of the medial prefrontal cortex, a region of the brain that had previously been implicated in mentalizing (i.e., inferring of other's beliefs, desires and intentions); however, no such activation occurred when people engaged with a computer. McCabe et al. [30] also replicated this pattern in the trust game. Riedl et al. [31] further replicated this result with virtual humans. Finally, in a seminal study, Sanfey et al. [35] showed that, when receiving unfair offers in the ultimatum game, people showed stronger activation of the bilateral anterior insula – a region associated with the experience of negative emotions – with humans, when compared to computers. In line with mind perception theories, this evidence suggests that people experienced less emotion and spent less effort inferring mental states with machines than with humans. This suggests that machines will fail to pass our benchmark, at least by default, in social decision making.

In digital games, Ravaja [14] demonstrated that people tend to show higher arousal and engagement with human than computer opponents. Specifically, people showed stronger EMG response in facial musculature (e.g., zygomaticus major), higher skin conductance, and better self-reported ratings with humans than computers. Moreover, the study showed that participants experienced stronger psychophysiological response with humans that were friends than strangers. This suggests that, in game-playing contexts, familiarity and long-term interaction may improve the likelihood that virtual agents will pass our benchmark.

Finally, research in social robotics tends to show that people behave differently with robots, when compared with humans. Kahn et al. [37] presented a study that clearly demonstrates this. In their experiment, children interacted for about 15 minutes with a humanoid robot, before an experimenter came into the room, interrupted the interaction, and asked the robot to "go wait in the closet". The question was whether this was fair to the robot, and whether the robot had any civil or moral rights. Effectively, children believed that the robot was entitled to fair treatment and had some rights; however, when compared to the case where this happened to an actual person, children were more likely to find the interruption unfair and to ascribe the person moral and civil rights. In line with mind perception theories, this results suggests that social robots will, thus, fail to pass our benchmark, at least by default.

3.2 Systems That Are As Good (or Better) Than Humans

Recently, de Melo, Carnevale, and Gratch [38] had participants engage in the ultimatum game with human or computer counterparts. The ultimatum game [39] is a simple

2-player game where there is a proposer and a responder. The proposer is given an initial endowment of money and then decides how much to offer to the responder. The responder then decides whether to accept or reject the offer. If the offer is rejected, no one gets anything. In this experiment, participants always assumed the role of proposers. The interesting aspect of the experiment, however, was that responders were manipulated to have different levels of mind. The experiment followed a $2 \times 2 \times 2$ factorial design: Responder (human vs. computer) × Agency (intentional vs. random) × Experience (non-emotional vs. emotional). For the manipulation of agency, they introduced a variation of the game where the responder was forced to make a random decision, independently of the offer, and the proposer was aware of this. For the manipulation of experience, the responder would either show a neutral facial display or show facial expressions of emotion. The emotion pattern rewarded fair behavior (e.g., sadness was shown when the offer was unfair or happiness when the offer was fair). The results showed, first, a main effect of Agency, with people offering more to intentional than random responders; nevertheless, there was no Responder × Agency interaction. The more interesting finding was a significant Responder × Experience interaction: when the responder showed no emotion, people offered more to human than computer responders, which is the usual bias in favor of humans; however, when responders showed emotion, people offered just as much to computers as they did to humans. Thus, adding appropriate emotion to computers was sufficient to "turn computers into humans", at least in the context of social decision making. These results show that perceptions of experience – the ability to sense and feel emotion – play an important role in making virtual humans pass our benchmark.

In a follow-up experiment, de Melo, Carnevale, and Gratch [40] demonstrated that it was possible to use other social mechanisms to overcome this intergroup bias people show in favor of humans. In particular, they explored multiple social categories [41]. This mechanism relies on the fact that people naturally categorize others as belonging to "in-groups", with which they identify with, and "out-groups". In their experiment, participants engaged with human or computer counterparts that were either of the same or different race as the participant (see Fig. 1). The results showed that, as usual, people offered more to humans than computers; however, people also made better offers to counterparts that shared the same race. In fact, there was no statistical difference between offers to computers of the same race and humans of a different race. The experiment, thus, showed that it is possible to use social categories – in particular, race – to help virtual humans pass the proposed benchmark.
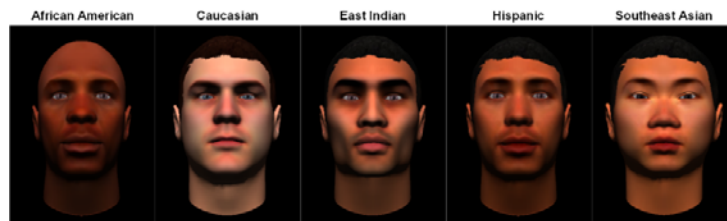
Fig. 1. People make more favorable offers to (human or non-human) counterparts that belong to the same social categories, such as race [40].

In yet another experiment, de Melo, Carnevale, and Gratch [40] demonstrated that multiple social categories could be used, not only to overcome but, to reverse people's bias in favor of humans. In this experiment, a third social category was created using the task payoff structure. In practice, this category created two teams. Participants were placed in the first team with two computers that shared the same race. In the other team, there were humans of a different race. So, in this case, computers were associated with two "positive" social categories (same team and race) and humans were associated with two "negative" categories (different team and race). As expected, people offered more to computers than to humans, thus, actually surpassing our benchmark.

3.3 Systems That Should Not Be Like Humans

In this subsection, we present three interaction contexts that seem to inherently favor machines to humans. First, we consider self-disclosure in health-screening interviews. In these clinical settings, it is important that patients disclose information about themselves honestly so that healthcare professionals may get an accurate medical history. In a recent study, Lucas et al. [42] demonstrated that when people believed that a virtual doctor was being controlled by algorithms, versus being driven by an actual person, people reported lower fear of self-disclosure, lower impression management, and were rated by observers as being more willing to disclose truthfully (Fig. 2).

Fig. 2. People are more willing to self-disclose honestly with virtual than real healthcare professionals.

Second, in the context of social robotics, Malle et al. [43] studied how people apply moral norms to robots, when compared to humans. They asked people how morally accepting was for a human or a robot to make an "utilitarian choice" in the trolley

dilemma. In this dilemma, a runway train is heading towards five workers in the tracks that will inevitably die, unless the decision maker, who is standing at a railroad intersection, pushes a lever that deviates the train away. However, the dilemma is that in the other track is a single worker, which will now be killed because the lever was pulled. Most people prefer to avoid making a decision, since they don't want to be responsible for the death of anyone. In this experiment, however, the results suggest that people would be more willing to accept the decision to pull the lever if it had been made by a robot, rather than by a human. Thus, if we assume that a decision needs to be made in such moral dilemmas, robots seem to be, by default, at an advantage when compared to humans.

Third, Sanfey et al. [35] showed that people were more willing to accept unfair offers in the ultimatum game if these were made by computers, rather than by humans. They further showed that this was happening because people experienced less negative emotion with computers than with humans. Therefore, if success is defined by the amount of money made, then it seems that computers are more likely to succeed than humans in making people accept unfair outcomes.

In all these social settings, it could be argued that people's decisions favor machines exactly because people have lower expectations of mental ability in machines. For instance, one might be more willing to accept unfair offers from a machine because a machine has no understanding of what it means to experience anger, or one might be more willing to self-disclose with a virtual human because one does not expect it to have the same kind of social concerns as humans (such as social image preservation). Thus, building on the mind perception framework, it would be interesting to confirm if, in these cases, proper simulation of mental ability in these machines – and emotional intelligence, in particular – would be sufficient to make people start treating them "just as badly" as they treat humans. Nevertheless, the main point here is that these systems should not aim to be like humans.

4   Conclusions

In this paper, we argued for a new benchmark for virtual agents that is more pragmatic, clearly defined, and useful than believability. The benchmark asks, in each specific social situation, that people behave with a virtual human in the same manner as with a real human. Thus, the benchmark serves as the basis for quantifying the difference between people's behavior with virtual and real humans. This benchmark can be easily measured in the lab, as demonstrated in the numerous studies reviewed in this paper. In fact, in many of these studies, the only thing that differed were participants' *beliefs* about whether they were interacting with a human or an autonomous agent. Moreover, the benchmark fits within a continuum, thus, allowing for continuous measurement of scientific progress towards the goal of achieving human-level social intelligence.

We also argue that perceptions of mind are critical for achieving virtual agents that are treated like humans in social settings. In particular, we reviewed evidence that perceptions of agency (the ability to plan and act) and experience (the ability to sense

and feel emotion) play a powerful effect on people's behavior with virtual agents. This research also emphasizes that people expect, by default, virtual agents to lack in experience and, therefore, appropriate simulation of emotional intelligence is especially important for passing our benchmark.

Our review also shows that there are social settings for which virtual agents seem to be inherently better than real humans, i.e., people tend to behave better – according to some domain-specific criterion – with virtual rather than real humans. As mentioned above, the mind perception framework suggests that appropriate simulation of mental ability in these cases, thus, may actually be detrimental to virtual agents. The point is that, in some cases, we do not want our systems to have the full gamut of capabilities that we see in humans. Future research should continue to study these social settings for which virtual agents that are unlike humans are particularly suited.

Finally, due to space restrictions, there were several topics that we chose not to address in the paper. First, in general, the focus was on virtual agents that attempt to perform social tasks that are usually expected of real humans. In this sense, we excluded agents that are meant to be different than humans by design (e.g., for entertainment purposes) or that are meant to serve as mere tools (e.g., a calculator). Second, our benchmark focused on behavioral realism; nevertheless, some researchers have emphasized that visual realism can also impact people's behavior (e.g., [44]) and, therefore, may warrant related, yet separate, benchmarks. Third, it is important to discuss how long-term interaction with virtual agents impacts people's behavior with them and, in particular, whether it facilitates or hinders achieving the proposed benchmark. Fourth, we avoided a discussion about the ethical issues associated with creating computers that behave just like real humans, given the different social and legal standing of artificially intelligent agents or robots (e.g., [45]). These are, nevertheless, important issues that need to be addressed as we quickly move towards a society that is surrounded by artificial agents that can match (and even surpass) the mental ability and social skill we see in humans.

Acknowledgments

References

1. Bates, J.: The role of emotion in believable agents. Comm. ACM 37, 122-125 (1994)
2. Mateas, M.: An Oz-centric review of interactive drama and believable agents. In: Wooldridge, M., Veloso, M. (eds.) Artificial intelligence today: Recent trends and developments, pp. 297-328, Springer-Verlag, Berlin (1999)
3. Riedl, M., Stern, A.: Believable agents and intelligent story adaptation for interactive storytelling. In: Proceedings of the Technologies for Interactive Digital Storytelling and Entertainment, pp. 1-12, Springer-Verlag, Berlin (2006)

9

4. Lester, J., Stone, B.: Increasing believability in animated pedagogical agents. In: Proceedings of the 1st International Conference on Autonomous Agents (AGENTS), pp. 16-21, ACM, New York (1997)

5. Rose, R., Scheutz, M., Schermerhorn, P.: Towards a conceptual and methodological framework for determining robot believability. Interac. St. 11, 314-335 (2010)

6. Riedl, M., Young, R.: An objective character believability evaluation procedure for multi-agent story generation systems. In: Proceedings of Intelligent Virtual Agents, pp. 278-291 (2005)

7. Reeves, B., Nass., C.: The media equation: How people treat computers, television, and new media like real people and places. Cambridge University Press, New York (1996)

8. Nass, C., Moon, Y.: Machines and mindlessness: Social responses to computers. J. Soc. Issues 56, 81-103 (2000)

9. Sundar, S., Nass, C. Source orientation in human-computer interaction: programmer, networker, or independent social actor? Commun. Res. 27, 683-703 (2000)

10. Nass, C., Moon, Y., Carney, P.: Are people polite to computers? Responses to computer-based interviewing systems. J. App. Soc. Psychol. 29, 1093-1109 (1999)

11. Nass, C., Fogg, B., Moon, Y. Can computers be teammates? Int. J. Hum.-Comput. Int. 45, 669-678 (1996)

12. Nass, C., Isbister, K., Lee, E.-J. Truth is beauty: Researching conversational agents. In: Cassell, J., Sullivan, J., Prevost, S., Churchill, E. (eds.) Embodied conversational agents, pp. 374-402, MIT Press, Cambridge, MA (2000)

13. Gajadhar, B., de Kort, Y., Ijsselsteijn, W.: Shared fun is doubled fun: Player enjoyment as a function of social setting. In: Proceedings of the Fun and Games Conference, pp. 106-117 (2008)

14. Ravaja, N.: The psychophysiology of digital gaming: The effect of a non co-located opponent. Media Psychol. 12, 268-294 (2009)

15. Hoyt, C., Blascovich, J., Swinth, K.: Social inhibition in immersive virtual environments. Presence 12, 183-195 (2003)

16. Okita, S., Bailenson, J., Schwartz, D.: The mere belief of social interaction improves learning. In: Proceedings of the Annual Meeting of the Cognitive Science Society (2007)

17. Weibel, D., Wissmath, B., Habegger, S., Steiner, Y., Groner, R.: Playing online games against computer- vs. human-controlled opponents: Effects on presence, flow, and enjoyment. Comp. Hum. Behav., 24, 2274-2291 (2008)

18. Katsyri, J., Hari, R., Ravaja, N., Nummenmaa, L.: The opponent matters: Elevated fMRI reward responses to winning against a human versus a computer opponent during interactive video game playing. Cerebral Cortex, Sep. 2012 (2012)

19. Lim, S., Reeves, B.: Computer agents versus avatars: Responses to interactive game characters controlled by a computer or other player. Int. J. Hum.-Comp. St., 68, 57-68 (2010)

20. Blascovich, J., Loomis, J., Beall, A., Swinth, K., Hoyt, L., Bailenson, J.: Immersive virtual environment technology as a methodological tool for social psychology. Psychol. Inq., 13, 103-124 (2002)

21. Blascovich, J., McCall, C.: Social influence in virtual environments. In: Dill, K. (ed.) The Oxford Handbook of Media Psychology, pp.305-315, Oxford University Press, New York (2013)

22. Epley, N., Waytz, A., Cacioppo, J. On seeing human: A three-factor theory of anthropomorphism. Psych. Rev. 114, 864-886 (2007)

23. Epley, N., Waytz, A. In: Fiske, S., Gilbert, D., Lindsay G. (eds.) The Handbook of Social Psychology, 5th edn, pp. 498-541, Wiley, New York (2010)

24. Waytz, A., Gray, K., Epley, N., Wegner, D.: Causes and consequences of mind perception. Trends Cogn. Sci., 14, 383-388 (2010)
25. Haslam, N. Dehumanization: An integrative review. Pers. Soc. Psychol. Rev. 10, 252-264 (2006)
26. Gray, H., Gray, K., Wegner, D. Dimensions of mind perception. Science 315, 619 (2007)
27. Loughnan, S., Haslam, N. Animals and androids: Implicit associations between social categories and nonhumans. Psychol. Sci. 18, 116-121 (2007)
28. Rilling, J., Sanfey, A.: The neuroscience of social decision-making. Annu. Rev. Psychol. 62, 23-48 (2011)
29. Gallagher, H., Anthony, J., Roepstorff, A., Frith, C.: Imaging the intentional stance in a competitive game. NeuroImage 16, 814-821 (2002)
30. McCabe, K., Houser, D., Ryan, L., Smith, V., Trouard, T.: A functional imaging study of cooperation in two-person reciprocal exchange. Proc. Nat. Acad. Sci. 98, 11832-11835 (2001)
31. Riedl, R., Moht, P., Kenning, P., Davis, F., Heekeren, H.: Trusting humans and avatars: Behavioral and neural evidence. In: Proceedings of the 32nd Int. Conference on Information Systems (2011)
32. Rilling, J., Gutman, D., Zeh, T., Pagnoni, G., Berns, G., Kilts, C.: A neural basis for social cooperation. Neuron 35, 395-405 (2002)
33. Krach, S., Hegel, F., Wrede, G., Binkofski, Kircher, T.: Can machines think? Interaction and perspective taking with robots investigated via fMRI. PLOS ONE 3, 1-11 (2008)
34. Kircher, T., Blumel, I., Marjoram, D., Lataster, T., Krabbendam, L, Weber, J. et al.: Online mentalising investigated with functional MRI. Neuroscience Letters 454, 176-181 (2009)
35. Sanfey, A., Rilling, J., Aronson, J., Nystrom, L., Cohen, J.: The neural basis of economic decision-making in the ultimatum game. Science 300, 1755-1758 (2003)
36. van't Wout, M., Kahn, R., Sanfey, A., Aleman, A.: Affective state and decision-making in the ultimatum game. Exp. Brain Res. 169, 564-568 (2006)
37. Kahn, P., Kanda, T., Ishiguro, H., Freier, N., Severson, R., Gill, B. et al.: "Robovie, You'll Have to Go into the Closet Now": Children's Social and Moral Relationships With a Humanoid Robot. Dev. Psychol. 48, 303-314 (2012)
38. de Melo, C., Carnevale, P., Gratch, J.: Bridging the gap between human and non-human decision makers. Presented at the Annual Meeting of the International Association for Conflict Management (2014)
39. Güth, W., Schmittberger, R., Schwarze, B. An experimental analysis of ultimatum bargaining. J. Econ. Behav. Organ. 3, 367-288 (1982)
40. de Melo, C., Carnevale, P., Gratch, J.: Social categorization and cooperation between humans and computers. In: Proceedings of the Annual Meeting of the Cognitive Science Society (2014)
41. Crisp, R., Hewstone, M.: Multiple social categorization. Adv. Exp. Soc. Psychol. 39, 163-254 (2007)
42. Lucas, G., Gratch, J., King, A., Morency, L.-P.: It's only a computer: Virtual humans increase willingness to disclose. Comp. Hum. Behav. 37, 94-100 (2014)
43. Malle, B., Scheutz. M, Arnold, T., Voiklis, J., Cusimano, C.: Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In: Proceedings of Human-Robot Interaction (2015)
44. Yee, N., Bailenson, J., Rickertsen, K.: A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces. In: Proceedings of CHI (2007)
45. Bingsjord, S.: Red-pill robots only, please. IEEE T Affect. Comput. 3, 394-397 (2012)