Computational Deception (Invited Talk)

Anton Nijholt

Abstract — In the future our daily life interactions with other people, with computers, robots and smart environments will be recorded and interpreted by computers or embedded intelligence in environments, furniture, robots, displays, and wearables. These sensors record our activities, our behaviour, and our interactions. Fusion of such information and reasoning about such information makes it possible, using computational models of human behaviour and activities, to provide context- and person-aware interpretations of human behaviour and activities, including determination of attitudes, moods, and emotions. Sensors include cameras, microphones, eye trackers, position and proximity sensors, tactile or smell sensors, et cetera. Sensors can be embedded in an environment, but they can also move around, for example, if they are part of a mobile social robot or if they are part of devices we carry around or are embedded in our clothes or body. Our daily life behaviour and daily life interactions are recorded and interpreted. How can we use such environments and how can such environments use us? Do we always want to cooperate with these environments; do these environments always want to cooperate with us? We argue that there are many reasons that human inhabitants of these environments do want to keep information about their intentions and their emotions hidden. Also their artificial interaction partner may have similar reasons to not give away all information they have or to treat their human partner as an opponent rather than someone that has to be supported. We survey situations where we can expect that human and artificial partner will not be honest to each other and will hide what they think, want or intend. These situations occur when the computer gets involved in social interaction, commerce and negotiation, and sports and games.

Index Terms — Conversations, Deception, Games, Human-computer interaction, Lies, Sports

1 Introduction

In 1982 Time Magazine made the computer 'Machine of the Year'. Until then, famous people such as Ronald Reagan, Lech Walesa and Ayatullah Khomeini had been chosen as 'Man of the Year'. That tradition was continued after 1982. Interestingly, in 2006 it was again the computer that appeared on the cover of Time magazine in its yearly election of the 'Person of the Year'. However, now it said, 'You.' 'Yes you. You control the Information Age. Welcome to your world.' This change from making the computer the machine of the year to a statement in which it is assumed to be necessary to make explicit that humans are in control illustrates that indeed, there can be doubts who is in control. Interaction with a computer in a human-like way has been the topic of research since the time of the early computers. Chatbots, question-answering systems and dialogue systems, have been designed and during the 1980s and the 1990s of the previous century such systems have been demonstrated in research environments. But, it is the computer that tells us how to issue commands and requests. The user has to adapt to the system, he or she is commanded to provide information at a time and in a way the system is assumed to be able to understand.

Despite slow progress in natural, intuitive and human-like human-computer interaction, it nevertheless remains a main research aim. There is optimism when looking at modelling human-computer interaction, in particular when looking at modelling nonverbal aspects of such interaction. New sensor technology has made it possible to track nonverbal interaction cues and activities. Current research activity, for example in various large-scale European research projects, is aiming at using sensor technology and sensor data interpretation of nonverbal aspects of human-human interaction and of human behaviour activity in general. Again, as has been the leading principle of research in the past, the assumption is that we can model human-human interaction, preferably in a multi-party interaction setting, and that this knowledge can be used to design more 'natural' interfaces between humans and computer-supported environments in 'daily-life' situations.

In these environments our daily life behaviour and daily life interactions are recorded and interpreted. How can we use such environments and how can such environments use us? Do we always want to cooperate with these environments; do these environments always want to cooperate with us? We argue that there are many reasons that users or rather human partners of these environments do want to keep information about their intentions and their

▪ *Anton Nijholt is with the Human Media Interaction Department of the University of Twente, Enschede, the Netherlands, E-mail: anijholt@cs.utwente.nl.*

emotions hidden from these smart environments. On the other hand, their artificial interaction partner may have similar reasons to not give away all information they have or to treat their human partner as an opponent rather than someone that has to be supported by smart technology. This will be elaborated in this talk. We will survey examples of human-computer interactions where there is not necessarily a goal to be explicit about intentions and feelings. Hence, we will look at (1) the computer as a conversational partner, (2) the computer as a butler or diary companion, (3) the computer as a teacher or a trainer, acting in a virtual training environment (a serious game), (4) sports applications (that are not necessarily different from serious game or education environments), and games and entertainment applications.

2 COOPERATION AND NON-COOPERATION

When modelling human-computer interaction, the main assumption is that users are cooperative. They have no choice. In [1] this is called the Cooperative Principle: Make your contribution such as it is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged. Speakers (generally) observe the cooperative principle, and listeners (generally) assume that speakers are observing it (conversational implicature). Grice also introduced some Conversational Maxims, such as the Maxim of Quantity: Make your contribution to the conversation as informative as necessary; and Do not make your contribution to the conversation more informative than necessary. There are other Maxims. The Maxim of Quality: Do not say what you believe to be false; and Do not say that for which you lack adequate evidence. The Maxim of Relevance: Be relevant (i.e., say things related to the current topic of the conversation). And, the Maxim of Manner: Avoid obscurity of expression; Avoid ambiguity; Be brief; and Be orderly.

This Gricean view has been attacked, extended and refined. There have been discussions about these principles and maxims and whether they are descriptive and prescriptive, refinements have been introduced that include ethical considerations and refinements have been introduced that look at the importance of nonverbal communication and how nonverbal signals should be included in these views (see e.g. [2]). Many researchers looked at ways to model mutual and cooperative understanding of face-to-face human-human understanding. And, they were quite willing to look beyond these Gricean principles.

As mentioned before, there are many daily-life situations where people don't follow the Gricean rules. That may be because of, for example, self interest, indifference, competition, politeness or respect. People may aim, consciously or unconsciously, at keeping the interaction going, rather than on aiming at satisfying the Gricean maxims. And, there are many real-life interaction situations that we want to model in human-computer or human-robot interaction where the computer, the social robot, or the virtual human may disagree with us, consider us an opponent rather than acting as a virtual agent that always agrees with us and performs our wishes, or simply tries to persuade us to do or forget about doing certain things. We can find such situations in electronic commerce applications, 'negotiation' situations, daily-life conversations, training and sports situations, and game and entertainment situations.

3 SITUATIONS REQUIRING NATURAL DECEPTIVE INTERACTIONS

3.1 Conversations, Q&A, & Dialogues

When we look at existing chatbots, question-answering systems, automatic speech telephone services or natural language dialogue systems it is clear that users have to adapt to the system and no natural interaction can take place. Users of these systems are assumed to be cooperative and in practice need to use a lot of empathy in order to get the system perform for what it has been designed.

As an extreme example consider the ALICE conversational system. Let's look at an example of an Alice dialogue that came up when we researched the possibility of a virtual diary companion for soldiers on a mission in Afghanistan. The virtual diary companion was designed to provide empathic support, stimulating soldiers to express their feelings while making entries in the diary [3].

Soldier: I've had a terrible day.
ALICE: What happened to it?
Soldier: First we were under attack.
ALICE: And then what happened? And second?
Soldier: Then our vehicle struck a landmine.
ALICE: Fascinating. It would be nice if our vehicle struck a landmine.
Soldier: Then my best friend got killed.
ALICE: Good reasoning. How well do you know this person?

It is assumed, and at least topic of research, that in the future there is the possibility to have dialogue and

conversational systems that will be used and will be appreciated by their users. This may be expected because the system learns to know about the user from his or her interaction history including preferences, interests, and attitudes. Moreover, apart from speech and language, there are other modalities that provide information from user to system. Gestures, facial expressions, body postures, gaze behaviour, and other information that can be measured (e.g., (neuro)physiological information) can be used by an interaction system to provide relevant feedback. It should be mentioned that in several human-human interaction studies it has been shown that there are situations where nonverbal cues are more important in conveying a speaker's message than the meaning of the words that are used. This is in particular true when attitudes or feelings are communicated. Cf. Mehrabian [4] who introduced the following rule: "Total Liking = 7% Verbal Liking + 38% Vocal Liking + 55% Facial Liking."

Consider the progress that is being made to make a computer a conversational partner. The computer can be represented as a virtual friend that knows about us and to whom we can talk in a confidential way. It may be a virtual butler that also knows about us and maybe even knows more about us than a real friend, to whom we can talk to in a less confidential way. In real life, whoever we talk to, we don't display all our feelings or are explicit about all our goals. We don't provide all information we have. Sometimes that is to protect ourselves from unwanted intimacy; sometimes it is to protect our conversational partner from information that may be harmful for him or her. We keep back information, we lie, and we manipulate. There are studies that tell us how many 'lies' we are using every day. In many cases these 'lies' are functional. They are not that important and they keep the conversation going. When we have useful applications for virtual conversational partners, do we always want them to be completely honest and really mean what they say during a conversation?

We can conclude that in social interaction settings, that is, in a setting where a virtual human or a social robot is used as a conversational partner, there are good reasons to have this artificial partner knowingly accepting that its human partner is not necessarily following the Gricean principles and adapts to a verbal and nonverbal exchange where it does not follow these rules itself (e.g. by displaying deceptive nonverbal behaviour [5], rather than what could be considered spontaneous behaviour). Detecting that a human conversational partner is not following these rules (consciously or unconsciously) is becoming possible by technology that senses all kinds of non-verbal communication information (from speech, gaze, head and body movements, and physiological information, including brain and muscle activity measurements...

3.2 Commerce, Negotiation, Persuasion

During daily conversations we are not always honest. These conversations do not necessarily have a particular aim. The situation is different when we consider verbal and nonverbal interaction between a human and an (embodied) agent in an electronic commerce setting. The aim of the agent, reflecting the aim of its designers and owners, can be to sell as many products and services as possible. Such an agent will not follow the Gricean cooperative principle. Neither will an agent that participates in an online auction or an agent that is meant to persuade a citizen to behave in a certain way. These agents take a certain perspective in their interaction and do not necessarily provide fully complete or fully correct information. They are not necessarily sincere. In their interaction with other agents or humans they have to decide when and how to honest and when and how to deceive and when and how to hide information. For that reason, in [6] it is argued that "Agents are and will be designed, selected or trained to deceive, and people will be deceived by and will deceive their own agents."

Even a personal assistant agent can decide to deceive its 'owner' or conceal certain information because it knows more about, among other things, legal consequences of actions, consequences for long-term goals and preferences that a user has, or consequences for health.

3.3 Teaching, Training, Serious Games

Computing intelligence and computing power can be embedded in a virtual teacher or a teaching environment. Teachers do not always act in an explicit cooperative way. It can be useful to provoke, challenge, or tease a student. It can be useful to use humour, to play the role of a non-understanding conversational partner and to display faked emotions. At the same time, a student interacting with a teacher or a virtual teacher has a strategy that aims at getting a good assessment of his or her knowledge and motivation. The computer-controlled virtual teacher needs to be aware of this. The student is not necessarily aware of the

strategies of a human teacher or the strategies that have been included in a virtual educational environment and an embodied virtual teacher. Neither the student, nor the teacher is playing according the Gricean rules. There is nothing wrong with that, but in order to act naturally and to be effective, a virtual teacher or teaching environment should be able to detect, analyze and synthesize such behaviour in order to generate understanding and empathic behaviour in order to take care of natural face-to-face interaction.

Virtual reality environments are used for teaching and training situations. In these environments events are simulated and trainees can 'enter' these environments in order to learn to collaborate with human or virtual team mates, enter into situations where they have to negotiate with human or virtual partners, or enter situations where they have to fight human or virtual opponents. An example of such a 'serious' or role-playing game is the virtual human doctor project [7]. The setting is a clinic somewhere in Iraq. The trainee is an army captain who has to persuade a doctor to move his clinic because of a planned military operation. This has to be done without revealing details of the military operation. Obviously, and being part of the training situation, the doctor is not necessarily cooperative. This requires the modelling of non-cooperative behaviour.

3.4 Sports, Games, and Entertainment

Presently we see research and the development of technology that aim at developing sensor-equipped and intelligent exercise and training environments. Microphones detect speech and sound, cameras detect movements of the body, the limbs and changes of facial expressions, there are sensors that detect positions and proximities, and physiological sensors provide information about body and mental state of a user of these exercise or training environments. These environments aim at improving the health of their users, for example by displaying a motivating virtual environment, a virtual coach, and a fitness exercise program. Interpretation of the information obtained from the sensors allows the environment, probably represented by a virtual human, to match actual behaviour with desired behaviour, and to adapt its appearance and its feedback strategies to the performance of the user [8]. One of the things that we noticed is that a trainer needs to be aware that a user is not necessarily honest in his or her verbal or nonverbal attitude towards a trainer. He or she can hide fatigue or exaggerate fatigue. The virtual trainer has the possibility to know about this and has to decide how to deal with this. This includes deciding whether the trainer's knowledge about a user's deceptive behaviour should be communicated to the user. It is not always in the interest of the user or the trainer to speak the truth.

In sports and games deceptive actions are part of the game. They are meant to divert attention from one's real purpose. Hence, in virtual training and recreational environments a trainer or in particular a game opponent is not only allowed, but also expected to have nonverbal behaviour that is aimed at deception. Just to mention an example, suppose we have a virtual fencing trainer. Its main job will be to exercise recognizing and generating deception behaviour. Similarly, we can look at virtual or mediated boxers, baseball players or rugby players [9].

4 MISTRUST

In many of the situations described above, the computer can interact with us in playful, exercise, entertainment, sports, and serious gaming environments. In these environments we can expect that situations we can expect that non-cooperative and deceiving behaviour is there. It is part of a game, it is part of training, and it is part of an exercise programme. It may be the case, and it was an essential theme in Stanley Kubrick's movie 2001, that we simply do not trust an advice or a decision made by an extremely intelligent computer and that we verbally and nonverbally try to deceive this computer, assuming that we know better. In a well-known fragment of this movie one of the astronauts (Dave Bowman) takes the decision to hide his suspicion that HAL, the intelligent computer, is not able to handle a particular dangerous situation. Or, at least, not willing to handle this situation in the (life-saving) interest of the astronauts. Dave decides to discuss this situation with his co-pilot, but is not aware that HAL has eyes everywhere and is aware of this discussion. Later, trying to convince HAL to adapt the mission's aims, HAL is able to confront Dave with this overheard discussion and refuses to make any changes to the mission. Nevertheless, Dave's empathy, trying to understand HAL's way of feeling and reasoning, turns out to be stronger than HAL's understanding of Dave's intentions. The '2001' movie is science-fiction, but nevertheless. The discussion between Dave and HAL is about trust, mistrust and

assuming that your conversational partner's aim has interests others than your and tries to deceive you.

> HAL: This mission is too important for me to allow you to jeopardize it.
>
> DAVE BOWMAN: I don't know what you're talking about, HAL?
>
> HAL: I know you and Frank were planning to disconnect me, and I'm afraid that's something I cannot allow to happen.
>
> DAVE BOWMAN: Where the hell did you get that idea, HAL?

HAL got that idea by observing Dave and Frank discussing how to deal with him while assuming their conversation was hidden from artificial eyes and ears sensors in the environment. Wrong idea, HAL knew.

Clearly, here, we cannot say what has to be done. Do we want to negotiate with the computer, do we want to compromise, or do we want to overrule the computer whatever his arguments are? Or is it up to the computer to choose among these alternatives? Whatever we choose, interaction models aware of different perspectives, different aims and different truths need to be designed.

5 CONCLUSIONS

There are many reasons why we need to deal with deceptive verbal and nonverbal interaction. We looked at natural conversations and why such conversations profit from not always showing true feelings, we looked at commerce and negotiation situations where users are not assumed to show their feelings, we looked at game, training and simulation environments where users have to compete, obey and adjust their behavior to demands and preferences of their coaches, their team mates, and their virtual opponents. In all these situations some modeling of non-cooperative behavior, empathic behavior, persuasive behavior, and some modeling of coaching or teaching behavior is required. In games we see research attempts to make the 'non-playing' characters more autonomous by providing them with intelligence and social behavior. When this is done, these characters need to know about competition, disagreement, aggressiveness and violence. There is also discussion about bringing games more into the real world, that is, it is expected that in the future more competitive situations will be designed in the real world that allow playful deception.

ACKNOWLEDGMENT

This research has been supported by the GATE project, funded by the Netherlands Organization for Scientific Research (NWO) and the Netherlands ICT Research and Innovation Authority (ICT Regie).

REFERENCES

[1] H.P. Grice: Logic and Conversation. In: Syntax and Semantics, Vol. 3, Speech Acts, ed. by P. Cole and J.L. Morgan. New York: Academic Press, pp. 41–58, 1975.

[2] J. Allwood. Linguistic Communication as Action and Cooperation. Ph.D. thesis, Göteborg University, Department of Linguistics, 1976.

[3] A. Nijholt, F. Meijerink, and P.-P. Maanen. A virtual diary companion. In: Fourth International Workshop on Human-Computer Conversation, Bellagio, Italy, pp. 1-5, 2008.

[4] A. Mehrabian. Silent Messages (1st ed.). Belmont, CA: Wadsworth, 1971.

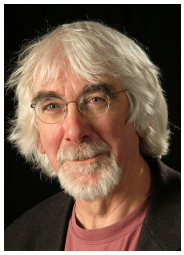
[5] I. Poggi, R. Niewiadomski, and C. Pelachaud. Facial Deception in Humans and ECAs. Modelling Communication. LNAI 4930, Berlin: Springer-Verlag, pp. 198-221, 2008.

[6] C. Castelfranchi. Artificial liars: Why computers will (necessarily) deceive us and each other? Ethics and Information Technology 2, Dordrecht: Kluwer Academic Publishers, pp. 113-119, 2000.

[7] D. Traum, W. Swartout, S. Marsella, and J. Gratch. Fight, Flight, or Negotiate: Believable Strategies for Conversing Under Crisis. Intelligent Virtual Agents (IVA 2005), LNCS 3661, Berlin: Springer-Verlag, pp. 52-64, 2005.

[8] Z. Ruttkay and H. van Welbergen. Elbows Higher! Performing, Observing and Correcting Exercises by a Virtual Trainer. In: Intelligent Virtual Agents. Lecture Notes in Computer Science 5208, Berlin: Springer-Verlag, pp. 409-416, 2008.

[9] S. Brault, B. Bideau, R. Kulpa, and C. Craig. How the global body displacement of a rugby player can be used to detect deceptive movement in 1 vs. 1. Proceedings of the 11th Virtual Reality International Conference, Laval-France, pp. 161-166, 2009.



Anton Nijholt studied mathematics and computer science at Delft University of Technology. His PhD research was done at the Vrije Universiteit in Amsterdam. After that he held several positions at universities in The Netherlands, Canada, and Belgium. He was research-fellow at the Netherlands Institute for Advanced Study in the Humanities and Social Sciences in 1995-1996. In 1989 he became full professor at the University of Twente in The Netherlands. Presently he is also scientific adviser of Philips Research. His research interests include multimodal interaction, natural language processing, virtual reality, embodied agents, entertainment computing, and brain-computer interfacing. Recently he has become interested in 'computational deception': How can we prevent that the computer knows too much about us?