When *Transparent* does not Mean *Explainable**

Extended Abstract[†]

Kerstin Fischer
Department of Design and
Communication
University of Southern Denmark
Denmark
kerstin@sdu.dk

ABSTRACT

Based on findings from interactional linguistics, I argue that transparency is not desirable in all cases, especially not in social human-robot interaction. Three reasons for limited use of transparency are discussed in more detail: 1) that social human-robot interaction always relies on some kind of illusion, which may be destroyed if people understand more about the robot's real capabilities; 2) that human interaction partners make use of inference-rich categories in order to inform each other about their capabilities, whereas these inferences are not applicable to robots; and 3) that in human interaction, people display only information about their highest capabilities, so that if robots display low-level capabilities, people will understand them as very basic. I therefore suggest not to aim for transparency or explainability, but to focus on the signaling of affordances instead.

CCS CONCEPTS

• **Human-centered computing** → HCI theory, concepts and models

KEYWORDS

Transparency, explainability, predictability of robot behavior

ACM Reference format:

K.Fischer. 2017. When 'transparent' does not mean 'explainable'. In *Proceedings of 'Explainable Robotic Systems'*, Workshop in conjunction with the HRI 2018 Conference, Chicago.

1 INTRODUCTION

Much recent work indicates the importance of transparency in human-robot interaction (e.g. Chen et al. 2014). Given that in interactions between people, interactants may undergo great efforts to understand who they are interacting with (e.g. Schegloff 1972; Clark 1998) in order to tailor their behaviors to their respective interaction partner (Schober & Clark 1989), it is expectable that people require no less information when interacting with computers (Fischer 2006) and robots (Fischer 2016). For instance, people ask questions about robots' capabilities ('What does it see?', 'Does he (sic!) see me?', 'Does it understand X?') (see Fischer 2016). Thus, there are similar needs for information about the interaction partner in human and in human-robot interactions, which suggests transparency to be of great importance.

Lyons (2013) argues that transparency is a term with many different possible meanings; for instance, in some studies, transparency is understood as users' understanding of why a robot behaved in an unexpected way (e.g. Kim & Hinds 2006), in others, it refers to information as a realistic judgment of the robot's reliability, such as the system's tendency for errors in a given context (e.g. Dzindolet et al. 2003), and in a third reading, it concerns how the robot communicates awareness of the human and his or her goals (e.g. Inagaki 2008).

However, not only is transparency a complex concept with many different facets, it may in fact not be equally desirable in all situations. An obvious case are human-robot interactions in Wizard-of-Oz scenarios (cf. Riek 2012). Here, it is a crucial aspect of the methodology that people do not understand how the robot is working. But also in other situations, a transparent robot is not necessarily an explainable robot because people carry specific expectations from linguistic interactions among people into human-robot interactions (cf. Hutchby 2001), which can lead to undesirable effects.

2 PROBLEMS WITH TRANSPARENCY

Previous work seems to assume that increased transparency necessarily means that the robot's behavior is more explainable and hence more predictable. In general, this may be true: For instance, if we know what a robot cannot do, we can make sure that we do the respective task ourselves; conversely, if we know that a robot carries out a particular task reliably, we do not need to monitor it all the time (see Lyons 2013). Thus, transparency can facilitate interaction. However, in social human-robot interactions, transparency may not always be suited to increase explainability or predictability because robots are artefacts and thus are man-made. This means that the robot's capabilities have been selected and implemented on the robot by a robot designer, and that they have not evolved as they do in humans. This has several consequences, which I discuss in the following.

2.1 Destroying the Illusion

One problem with transparency is that robot functionalities may be so basic or so different from human capabilities that knowledge about how the robot really functions would be off-putting and disillusioning. Work in HRI has shown ubiquitously that people anthropomorphize robots such that they treat them as if they had more human-like capabilities, behaviors and intentions than they really have (e.g. Fussel et al. 2008). While one can dismiss this behavior as irrational (e.g. Nass & Moon 2000), anthropomorphic responses to robots are often desired; in fact, there would be no social HRI if people did not anthropomorphize robots. Thus, in these cases, the robot should not reveal too much about its real functions.

2.2 Isolated Capabilities

Since robot capabilities have not evolved, more complex behaviors do not implicate less complex behaviors, as they would in humans (Fischer & Moratz 2001). For instance, in human interaction, people may elicit information about their interaction partners that is information-rich (Sacks 1984); for example, they may ask about their interaction partner's occupation because this allows them to infer the kind of vocabulary the partner may be familiar with (Clark 1998). In contrast, regarding robots, such inferences cannot be drawn due to the selected capabilities the robot was implemented with by the robot designer. Fischer & Moratz (2001) report about an experiment in which the robot designer had implemented very high-level spatial language capabilities into the robot, but not more basic ones. If for some reason participants encountered an error, they only tried more basic capabilities, and if they did not succeed with this, they made their instructions even simpler. Not a single participant chose a more complex instruction strategy after encountering an error. Thus, for humans it is inconceivable that someone can have higher-level capabilities without having gone through the more basic levels first (Fischer 2006). For transparency this means that signaling the robot's capabilities may be misleading because people will infer that the robot has all the basic capabilities that humans assume to be preconditions for the higher capability to develop.

2.3 Downwards Evidence

One may argue that specifically because people will make inferences on some robot capabilities based on other capabilities the robot needs to make all of its capabilities transparent, high-level and low-level, thus rendering inferencing superfluous. However, if a robot signals that it has a certain low-level capability, people will infer that it does not have *any* higher level capabilities; because people in human interaction signal to each other only inference-rich categories, they generally make use of downwards evidence (Clark 1996). For instance, when giving feedback, people may signal only that they agree to the partner's utterance, which provides downwards evidence that they have also heard and understood the utterance. If they signal that they have understood the partner, this is taken to mean that they do not agree, because otherwise they would have said so (cf. Allwood et al. 1992). Similarly, when a robot signals that it has understood a word correctly, the implication is that it has in fact problems with such basic tasks, which leads users to dismiss the robot as too basic (Fischer 2011). Thus, being transparent about low-level capabilities leads to unwanted inferences about the robot's capabilities.

3 DISCUSSION

The examples discussed indicate that especially in social human-robot interactions, revealing the robot's real capabilities and underlying processing may hinder, rather than improve, interaction; this is on the one hand due to the fact that interactions with social robots necessarily rely on some kind of pretense, during which robots' 'social' capabilities are being simulated (cf. Seibt 2017). On the other hand, transparency may create unwanted effects because it is especially the implicitness of information that makes human interactions so smooth and seamless. People transfer their interactional practices to interactions with robots since they are so common that they do not generally notice them themselves (cf. Hutchby 2001).

Moreover, the discussion above suggests that there can be a trade-off between transparency and simulation: Especially in social human-robot interactions, knowledge about the robot's capabilities may destroy the illusion necessary for social interaction. If possible, signaling how robots function may not be the best option.

4 DESIGN RECOMMENDATIONS

So how can robots be made explainable if transparency is not always the best strategy? I would like to make the following five recommendations:

- design the robot with explainability in mind
- exploit ambiguities in pragmatic function
- use downwards evidence wherever possible
- employ inference-rich categories
- make the robot's affordances visible

In particular, one possible solution from a systems design perspective is to make robots more explainable in those situations in which people expect a particular hierarchy of complexity of robot capabilities by anticipating people's expectations about what basic capabilities underlie the more advanced capabilities they endow their robots with. By developing not isolated robot behaviors but clusters of basic and advanced behaviors, system designers can avoid many interactional problems.

Another conclusion we can draw from the above is that only the most complex capabilities should be signaled, and that signaling of failure on very low-level capabilities should be avoided. One possibility to address this issue is to exploit natural ambiguities of human interaction. For instance, the question 'are you ready?' can be asked for politeness reasons, providing the communication partner with the opportunity to determine the next step, and equally plausibly because the speaker really does not know whether the communication partner has finished doing what he or she was doing. This ambiguity can be exploited in those cases in which the robot does not have access to the human's actions, without revealing this shortcoming directly.

Furthermore, provided that system design has indeed implemented also the respective basic capabilities on which a particular robot capability would rest if the robot had evolved, the natural inferencing of human interaction can be exploited by making use of downward evidence; that is, the robot could produce signals on the highest level of complexity and take these to include the low-level ones. One such example is the use of relevant next utterances; instead of signaling directly what was understood, the robot may present a relevant next utterance, such as a clarification question (cf. Fischer 2016).

Most crucially, however, my suggestion would be to make use of natural human negotiation processes by means of which they provide each other with inference-rich categories (see Clark 1998) and to design for signaling the affordances of the robot, i.e. how it is to be used. For instance, when human interactants bring up their professional identities, relationship with the host, number of children or other such small talk topics in conversation, then they implicitly provide their communication partners with information about how they should be talked to, helping their partners to tailor their behavior for them. In human-robot interaction, experiment participants often ask very similar questions about the robot, or address these questions to the robots themselves, such as what the robots see, what language they understand, what they can do and how they should be interacted with. Many of these functionalities can be signaled by means of the robot design, i.e. providing the robot with obvious visual and acoustic sensors, a touchpad, arms, wheels etc., but others can be signaled by means of verbal behavior, for instance, by means of 'leading' questions such as 'to which place do you want me to go?' Such a question prevents most people from saying 'go 80cm straight, then make a 45 degree turn' or 'slightly left'. Thus, interaction design could focus on indicating the robot's affordances implicitly, i.e. on giving indications to the users as to how to interact with the robot, rather than signaling its capabilities.

ACKNOWLEDGMENTS

This work was partially supported by the Danish Innovation Foundation in the context of the *Smooth* Project.

REFERENCES

[1] Allwood, J., J. Nivre, and E. Ahls´en (1992). On the semantics and pragmatics of linguistic feedback. Journal of Semantics 9, 1–26.

[2] Chen, T., Campbell, D., Gonzalez, F., and Coppin, G. (2014). The effect of autonomy transparency in human-robot interactions. In Proceedings of Australasian Conference on Robotics and Automation, Melbourne.

[3] Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

[4] Clark, H. H. (1998). Communal lexicons. In K. Malmkjaer and J. Williams (Eds.), Context in Language Learning and Language Understanding, pp. 63–87. Cambridge: Cambridge University Press.

[5] Dzindolet, M., Peterson, S., Pomranky, R., Pierce, L., and Beck, H. (2003). The role of trust in automation reliance. *Journal of Human-Computer Studies*, 58:697–718.

[6] Fischer, K. (2006): What Computer Talk Is and Isn't. Human-Computer Conversation as Intercultural Communication. Linguistics – Computational Linguistics 17, Saarbrücken: AQ.

[7] Fischer, K. (2011): How people talk with robots – designing dialog to reduce users' uncertainty. *AI Magazine* 32 (4): 31-38.

[8] Fischer, K. (2016): Designing Speech for a Recipient: The Roles of Partner Modeling, Alignment and Feedback in So-Called 'Simplified Registers'. Amsterdam: John Benjamins.

[9] Fischer, K. & Moratz, R. (2001): From Communicative Strategies to Cognitive Modelling. *Proceedings of the First International Workshop on 'Epigenetic Robotics'*, Lund, Sweden.

[10] Fussel, S.R, Kiessler, S., Setlock, L.D. and Yew, V. (2008). How People Anthropomorphize Robots. *Proceedings of HRI'08*, Amsterdam, p. 145-152.

[11] Hutchby, I. (2001). Conversation and Technology: From the Telephone to the Internet. Cambridge: Polity.

[12] Inagaki, T. (2008). Smart collaboration between humans and machines based on mutual understanding. Annual Reviews in Control. 253-261.

[13] Kim, T. and Hinds, P. (2006). Who should I blame? effects of autonomy and transparency on attributions in human-robot interactions.

[14] Lyons, J. B. (2013). Being transparent about transparency: A model for human-robot interaction. In 2013 AAAI Spring Symposium Series.

[15] Riek, L. (2012). Wizard of Oz studies in HRI: A systematic review and new reporting guidelines. Journal of Human-Robot Interaction, 1(1).

[16] Sacks, H. (1984). Notes on methodology. In J. Atkinson and J. Heritage (Eds.), Structure of Social Action: Studies in Conversation Analysis, pp. 21–27. Cambridge: Cambridge University Press.

[17] Schegloff, E. A. (1972). Notes on a conversational practice: Formulating place. In D. Sudnow (Ed.), Studies in Social Interaction , pp. 75–119. New York: Free Press.

[18] Schober, M. and H. H. Clark (1989). Understanding by addressees and overhearers. Cognitive Psychology 21, 211–232.

[19] Seibt, J. 2017. Towards an Ontology of Simulated Social Interaction: Varieties of the "As If" for Robots and Humans. In: R. Hakli, J. Seibt (eds.), *Sociality and Normativity for Robots*, Studies in the Philosophy of Sociality 9. Springer.