Embodiment is a Double-Edged Sword in Human-Robot Interaction: Ascribed vs. Intrinsic Intentionality

Tom Ziemke
iLab, School of Informatics
University of Skövde, Sweden
& HCS, IDA
Linköping University, Sweden
+46-705-441444
tom.ziemke@his.se

Serge Thill
Interaction Lab
School of Informatics
University of Skövde
54128 Skövde, Sweden
+46-500-448389
serge.thill@his.se

David Vernon
Interaction Lab
School of Informatics
University of Skövde
54128 Skövde, Sweden
+46-500-448392
david.vernon@his.se

ABSTRACT
This very short paper makes a relatively simple point: The human embodied cognitive capacity / tendency to attribute intentionality, goals, etc. to others, and to interpret their behavior in intentional terms, is fundamental to many types of social interaction. There are at least two quite different conceptions of embodied cognition though, underlying much research in cognitive robotics and human-robot interaction, which also differ regarding whether robots (a) could actually have their 'own' intrinsic intentionality, or (b) could only be ascribed/attributed intentionality, similar to the way cartoon characters are. For robotics research as such the distinction might be secondary, and for philosophy of mind the questions might not be resolvable any time soon. For society and the general public, however, the issue potentially has quite significant social and ethical implications – therefore researchers might need to pay more attention to this than they have so far.

Categories and Subject Descriptors
I.2.0 [**Computing Methodologies**]: Artificial Intelligence – *Philosophical foundations.*

General Terms
Human Factors, Theory.

Keywords
Embodied cognition, social interaction, intentionality, autonomy.

1. INTRODUCTION
As Sciutti and colleagues recently pointed out, the "ability to understand others' actions and to attribute them mental states and intentionality is crucial for the development of a theory of mind and of the ability to interact and collaborate" [1]. While this is central to human embodied social interaction, it is clear that the underlying cognitive capacity is not limited to interpreting the behavior of other people. It is well known from the classic studies of Heider and Simmel [2] that humans also tend describe the behavior of simple moving objects (triangles, circles, etc.) in intentional terms. Naturally, this also applies to more complex and human-like objects, such as cartoon characters. When, for example, we see Donald Duck angrily chasing chipmunks Chip and Dale because they are stealing his popcorn, it comes very natural to us to interpret their behavior in intentional terms – as illustrated by the first half of this sentence. At the same time, however, we presumably all understand that Donald, Chip, and Dale are not real and therefore also do not really have intentions.

Not surprisingly, this attribution also extends to different types of technology, in particular more or less autonomous systems. In the case of autonomous vehicles, recent research [3] indicates that anthropomorphism – "a process of inductive inference whereby people attribute to nonhumans distinctively human characteristics, particularly the capacity for rational thought (agency) and conscious feeling (experience)" – increases the trust people have in such systems. Hence, it is very likely that in social interactions with robots, humanoid ones in particular, (a) humans will attribute agency, intentionality, etc. to such artifacts (cf. [1]), and (b) interactions benefit from such attributions.

This, however, also raises the question what exactly is the status of the intentions, goals, etc. that we ascribe to robots? After all, they are real (physical), they are interactive, and in the humanoid case they behave and look human-like to some degree. Does that mean that, like humans, they potentially have their own intrinsic intentionality? Or is their intentionality, as in the case of Donald Duck, necessarily only ascribed?

2. DISCUSSION
Researchers in AI and robotics typically avoid answering this question explicitly and might prefer to dismiss it as a purely philosophical question. Implicitly, however, research in embodied AI and robotics, touches on the issue quite commonly. For example, embodied AI researchers commonly refer to Searle's 1980 *Chinese Room Argument* [4] to illustrate that traditional – *disembodied* – approaches to AI were deeply flawed because they only dealt with the internal manipulation of representations by computer programs. Understanding the details of the argument is not relevant to this short paper, but in a nutshell, Searle's criticism was that "the operation of such a machine is defined solely in terms of computational processes over formally defined elements", and that such "formal properties are not by themselves constitutive of intentionality" [4]. Researchers in embodied AI/robotics commonly argue that these problems of traditional AI can be overcome by *'embodied'* (robotic) approaches to AI, which allow internal representations/mechanisms to be grounded in sensorimotor interactions with the physical and social environment. This, however, completely ignores the fact that already back in 1980, in the original paper, Searle presented – and rejected – what he called the *'robot reply'*, which entailed pretty much exactly what is now called *embodied AI*, i.e. computer programs running on robots that interact with their environment.

What is interesting in this context is that there are quite many researchers who – like Searle – take the Chinese Room Argument to be a valid argument against traditional AI, but at the same time

– unlike Searle – consider the *physical/sensorimotor embodiment* provided by today's robots to be sufficient to overcome the problem. In Harnad's terms, this type of embodied AI has gone from a *computational functionalism* to a *robotic functionalism* [5]. Zlatev, for example, formulated the latter position very explicitly, arguing that there is "no good reason to assume that intentionality is an exclusively biological property (pace e.g. Searle)", and "thus a robot with bodily structures, interaction patterns and development similar to those of human beings … could possibly recapitulate [human] ontogenesis, leading to the emergence of intentionality" [6]. Others, including Searle naturally, do indeed believe that there are good reasons to assume that intentionality is in fact a biological property intrinsic to living bodies [7, 8, 9].

This illustrates Chemero's point that there currently are (at least) two very different positions that are both referred to as *'embodied cognitive science'* [10]. The one that Chemero refers to as *radical embodied cognitive science* is grounded in anti-representationalist and anti-computationalist traditions. The other, more mainstream version, on the other hand, in line with robotic functionalism, is derived from more or less traditional representationalist and computationalist theoretical frameworks.

It is therefore not surprising that researchers in embodied AI and robotics, inspired by theories of embodied cognition, are confused regarding the relevance of 'embodiment', and subsequently find it difficult to answer the question to what degree their robots have, or could have, their own intrinsic intentionality. If you adopt the more mainstream position of robotic functionalism, then, as Zlatev put it, there is "no good reason to assume that intentionality is an exclusively biological property" [6]. If, on the other hand, you adopt a more radical, non-functionalist position, e.g. the enactive view, which has also gained some influence in embodied AI [8, 9, 11, 12], then intrinsic intentionality indeed might very well be "an exclusively biological property" and therefore most probably not replicable in robots with current technology.

If at this point you are about to dismiss (once more) the issue of robot intentionality as a purely philosophical question – which obviously the philosophers cannot answer either – it should be noted that the context of human-robot interaction adds a novel dimension to this old problem and gives new social and ethical relevance. The point is that human interaction with robots is to some degree comparable to interaction with animals: It is certainly the case that scientists, philosophers, and the general public are divided regarding whether or not animals have and experience human-like feelings. However, neither the fact that we cannot conclusively answer that question, nor the fact that we are likely to have different opinions (e.g., vegans vs. vegetarians vs. meat-eaters), change the fact that we need to have legislation, ethical guidelines, personal positions, etc. for how animals are to be treated in our society. Likewise, in the case of robots, whether or not we want to deal with the philosophical problem of robot intentionality, if or when robots become a part of human society, we will need to come to some kind of conclusion anyway.

3. SUMMARY AND CONCLUSION

To summarize, the proverbial double-edged sword mentioned in the title is this: The human cognitive capacity/tendency to interpret behavior as intentional is central to embodied social interactions and to our way of interpreting both the animate and the inanimate world. For human-robot social interaction, this capacity is likely to be very useful because it tends to "fill in the gaps" and make interactions more natural and trustworthy. However, the tendency to attribute human-like mental states also comes with the tendency to view things as more human-like than they maybe really are. For cartoon characters this is obvious, and most people have no problems at all to understand that Donald chases Chip and Dale because he is angry and wants his popcorn back, and at the same time understand that neither Donald nor the chipmunks are real, and therefore none of them eats or wants popcorn anyway. For robots, this is much less obvious, because unlike cartoon characters, they are real, physical, 'embodied', etc., they are physically and socially interactive, and in the humanoid case they often also look and behave human-like to some degree.

The question therefore is if their apparent intentionality is only ascribed by the observer, as in the case of cartoon characters, or in fact is genuine and intrinsic to those robots themselves. Your answer to the question is likely to depend on your conception of embodied cognition – and the body underlying it. If you adopt the current mainstream position of robotic functionalism, according to which intentionality arises from the physical body's sensorimotor interaction with the environment, then intrinsic intentionality in robots is at least possible. If, on the other hand, you adopt the view that embodied cognition is ultimately grounded in the living body, then the intentionality of at least current-technology robots is necessarily only ascribed. Which of these positions we, as a society, adopt in the future is likely to have significant social and ethical consequences for the way we deal with robots.

4. ACKNOWLEDGMENTS

Supported by the European Commission, FP7 project 611391, DREAM (*Development of robot-enhanced therapy for children with autism spectrum disorders*), and the Knowledge Foundation, SIDUS project AIR/TINA (*Action and intention recognition in human interaction with autonomous systems*).

5. REFERENCES

[1] Sciutti, A., Bisio, A., Nori, F., Metta, G., Fadiga, L., and Sandini, G. 2014. Robots can be perceived as goal-oriented agents. *Interaction Studies*. 14, 3, 329-350.

[2] Heider, F., Simmel, M., 1944. An experimental study of apparent behavior. *American J. of Psychology* 57, 243–259.

[3] Waytz, A., Heafner, J., & Epley, N. 2014. The mind in the machine. *J. of Exp. Soc. Psychology* 52*,* 113-117.

[4] Searle, J. 1980. Minds, brains, and programs. *Behavioral and brain sciences* 3, 3 , 417-424.

[5] Harnad, S. 1989. Minds, machines and Searle. *Journal of Experimental & Theoretical Artificial Intelligence* 1, 1, 5-25.

[6] Zlatev, J. 2001. The epigenesis of meaning in human beings, and possibly robots, *Minds and Machines* 11, 2, 155-195.

[7] Varela, F. 1997. Patterns of Life: Intertwining Identity and Cognition. *Brain & Cognition* 34, 72-87.

[8] Ziemke, T. 2008. On the role of emotion in biological and robotic autonomy. *BioSystems* 91, 401-408.

[9] Froese, T., and Ziemke, T. 2009. Enactive artificial intelligence. *Artificial Intelligence* 173, 466-500.

[10] Chemero, T. 2009. *Radical embodied cognitive science*. MIT Press, Cambridge, MA.

[11] Vernon, D. 2010. Enaction as a conceptual framework for developmental cognitive robotics. *Paladyn* 1, 2, 89-98.

[12] Vernon, D. 2014. *Artificial cognitive systems: A primer*. MIT Press, Cambridge, MA.