Reflections of Consciousness: The Mirror Test

Pentti O A Haikonen

Nokia Research Center
P.O. Box 407
FI-00045 NOKIA GROUP
Finland
pentti.haikonen@nokia.com

Abstract

Humans and some animals are able to recognize themselves in a mirror. This ability has been taken as a demonstration of self-consciousness. Consequently, it has been proposed that the self-recognition in the mirror image, the mirror test, could also be used to determine the potential self-consciousness of cognitive machines.

It is shown that very simple machinery is able to pass the mirror test and consequently it is argued that the passing of the mirror test per se does not demonstrate the existence of self-consciousness. Next, the realization of more complex machinery that passes the mirror test and could be considered to possess self-consciousness is outlined.

Introduction

Humans and some animals recognize their mirror image while other animals may treat their mirror image as another animal behind the mirror. Human babies under 18 months of age do not normally recognize their mirror image. Psychological tests have shown that babies seem to become socially self-aware around the time when they begin to recognize themselves in a mirror (Lewis et al 1989). This observation would seem to link self-awareness and mirror image recognition together. Consequently, mirror image recognition, especially the so-called rouge test, has been used in developmental psychology to determine the time of the emergence of self-awareness in very young children. In the seventies Gordon Gallup applied the mirror test to chimpanzees, who passed the test (Gallup 1970). Most other animals including cats and dogs do not pass this test as many of us may have personally noticed.

It has been proposed that the mirror test could also be applied to potentially conscious machines; if the machine is able to recognize itself in the mirror then the machine should be self-aware. Junichi Takeno of Meiji University, Japan, has built small robots that demonstrate different inner activity when confronted by their mirror image and another similar looking robot. Takeno proposes that this is a demonstration of the robot's self-awareness (Takeno, Inaba and Suzuki 2005). Michel, Gold and Scassellati of Yale University have programmed their robot "Nico" to recognize the mirror image of its hand by the simultaneity of the hand motion command and the visually detected motion of the mirror image (Michel, Gold and Scassellati 2004, Gold and Scassellati 2005, 2006). Fitzpatrick, Arsenio and Torres-Jara at MIT have implemented similar binding between the proprioceptive information and visual mirror image information of robot hand movement in the Cog robot for the purpose of mirror image self-recognition (Fitzpatrick and Arsenio 2004, Fitzpatrick, Arsenio and Torres-Jara 2004). Superficially these robots might seem to be more self-aware than most animals; however, the situation may be more complex than that. Self-awareness and mirror image recognition may go together, but the lack of the latter may not necessarily prove the lack of the former. (Are cats definitely without any kind of self-awareness?) Also, the presence of the latter may not necessarily be a proof of the former as will be shown in the following.

Mirror Image Recognition Mechanisms

How does one learn to recognize one's own face in a mirror? Obviously no visual pattern matching can be done, as initially one does not have any visual reference image of one's appearance. Therefore the mirror image self-recognition cannot be based on specific features within the visual modality only. However, one may utilize certain strategies that may lead to the mirror image self-recognition.

Firstly, there are body parts that can be seen directly and this visual information may be compared to the mirror image of the same parts. In this way, by the common visual features and common motion the correspondence and causal connection between the body part and its mirror image may be discovered. Thus, for instance, the mirror image of a hand may be taken to represent the real hand, even though this conclusion is not necessarily inevitable. This outcome would allow the idea of a mirror; things seen in a mirror represent real objects that are outside the mirror. Thus, when one's hand touches the face, the mirror image of the hand is known to represent one's own hand and consequently the mirror image of the face may also be taken to represent one's own face and not something that is behind the mirror.

Secondly, facial mirror image self-recognition may arise from the utilization of amodal features that are related to facial expressions. Amodal features are features that are shared by several sensory modalities and manifest themselves simultaneously. Facial expressions are generated by the commanding of certain facial muscles. The activity of these muscles lead to the excitation of cutaneous receptors in the skin, which in turn leads to corresponding somatosensory sensations. The timing of the somatosensory sensation patterns corresponds to the perceived timing of the visual change in the mirror image and is in this case the amodal feature that can be utilized. On the other hand the somatosensory sensations relate to the bodily self and the coincident changes in the mirror image may lead to the association of the somatosensory bodily self with the mirror image. In the following this approach is elaborated.

Trivial Passing of the Mirror Test

The principle of the utilization of amodal features in mirror image self-recognition can be easily demonstrated by trivial electronic circuits. As an example a simple case is considered. In this example the "face" of the robot contains one blinking light emitting diode LED for "facial expressions" and one photodiode as an "eye". If the reflected light from the system's own "facial" LED and the light from a similar external blinking LED cause different output states then the system can be said to pass the mirror test, see Figure 1.

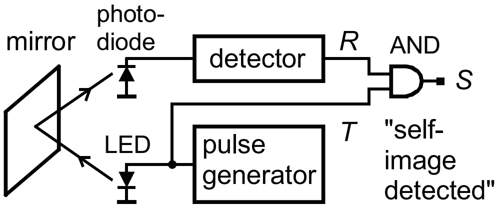

Figure 1. A trivial circuit that "passes" the mirror test

In Figure 1 a pulse generator supplies a train of digital electric pulses *T* to the "facial" light emitting diode LED causing it to blink. The blinking light from that LED is reflected via a mirror to the "eye", a photodiode. The photodiode receives the blinking light and the detector circuit generates the corresponding train of received digital pulses *R*. If these pulses are caused by the blinking light from the "facial" LED then the signals *R* and *T* correspond to each other and appear simultaneously. This correlation can be detected by the logical AND function: *S* = *R* AND *T*. If *S* = 1 then the mirror image of the self-generated light pulse is detected. If, on the other hand, the photodiode does not receive any light pulses or receives light pulses from some external sources then no correspondence between the signals *R* and *T* exist and the *R* AND *T* function gets the value zero. Thus the circuit behaves differently when it receives blinking light from external sources or receives the mirror image of its own blinking led. (In practice some kind of temporal filtering would be required to filter out transient random correlations.)

This kind of circuitry, perhaps with infrared diodes that conceal the true nature of operation, could be easily built into a doll. The output of this circuit might be used to trigger suitable hand motions and perhaps stored voice messages like "It's me", whenever the doll were positioned in front of a mirror. To an external observer the doll would seem to be able to recognize itself in the mirror. In reality though, this doll would not be conscious at all.

The circuit of Figure 1 is a most trivial one, but it should be obvious that the principle would apply also to more complicated situations. Instead of a solitary LED there could be an array of LEDs depicting a face. Or, there could be an artificial face that could display various expressions effected by corresponding motor commands. The self-image detection principle of Figure 1 would still work, as the detection process is based on amodal properties: the timing and temporal pattern of the motor commands and the received visual changes. It is also obvious that various artificial neural network schemes could be used to execute the correlation AND function.

This example should show that it is easy to build machines that "recognize" themselves in a mirror. However, the author argues that this kind of recognition is not true self-recognition because these machines do not have an inner concept of self or a general "self-model" or a "sense of self", which, for instance, Holland and Goodman (2003) and Kawamura et al (2005) find necessary. These machines would not possess any concept of self that could be evoked by the "self-image detected" signal. Therefore this kind of passing of the mirror test is trivial and does not have much to do with consciousness.

Somatosensory Grounding of Self-Concepts

Humans and robots have bodies that remain the same when the environment changes. Wherever I go, my body goes, too, and the sensations of my body follow me while the percepts of environment change. Therefore the concept "I" is naturally associated with the body. The brain, however, cannot associate concepts and objects with each other directly, only internal representations of these can be associated. Thus, for example, Damasio (2000, p. 22) has proposed that at the basic level the sense of self is grounded to the representations (percepts) of the body. This view is shared by the author.

The boundary of the body is the skin, which is saturated with various receptors for touch, temperature and pain (cutaneous sensations). Touch information is provided by mechanoreceptors that are sensitive to local pressure, vibration and stretching. The position and motion of the limbs are sensed by stretch receptors in the muscles and joints (kinesthetic sensing). The subsystem in the brain that handles these own-body related sensations is the somatosensory system. Obviously, the somatosensory system deals with information that is necessary for the generation of a bodily self-image.

A robot that is supposed to have a bodily self-image should also be able to acquire the same own-body information and have the equivalent of the somatosensory system. This means that the robot should have the equivalent of the skin and its cutaneous receptors as well as sensors that correspond to the kinesthetic receptors.

Passing the Mirror Test Properly

A useful concept of self would allow "a body model" and the evocation of the "imagery" of the body parts and their locations when necessary. This in turn would allow the touching of these parts, if physically possible. A concept of self should also allow the naming of the self and the utilization of the "self" as an executing agent in imagination and planning. The concept of self should also include the self's personal history because the planning of future actions could not be done properly if the system could not remember what has already happened.

Successful passing of the mirror test would not only contain different inner activity for the percepts of the mirror image and the percepts of similar foreign objects. Instead, the recognition of the mirror image should lead to meaningful consequences, which could be demonstrated e.g. by the evocation of one's name and the ability to control movement by the help of the mirror image.

A simplified cognitive architecture for an embodied robot along the principles of Haikonen (2003, in more detail 2007b) is depicted in Figure 2.

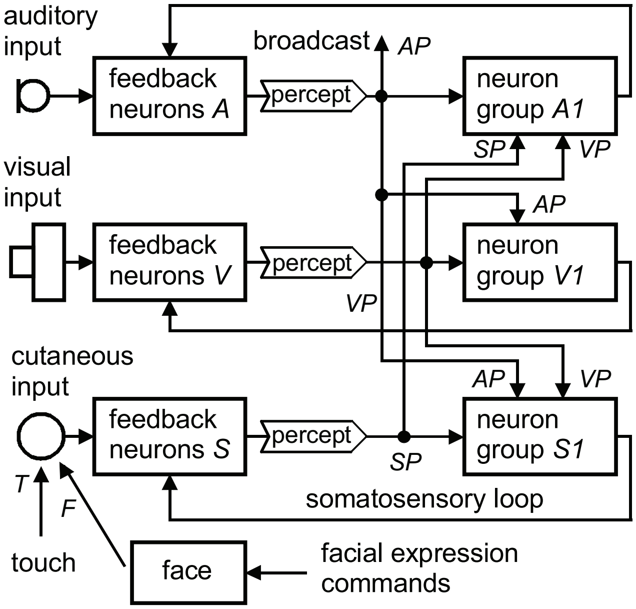

Figure 2. A cognitive architecture with somatosensory self-concept

In Figure 2 perception/response feedback loops for auditory, visual and cutaneous sensory modalities are shown. The cutaneous module is a part of the somatosensory system and, in addition to its basic function, the perception of cutaneous information, grounds self-related percepts to somatosensory percepts.

In Figure 2 the signal *SP* of the somatosensory loop is to be understood as a generalized signal that will arise whenever somatosensory percepts arise (logical OR-function of all possible somatosensory percepts). Thus, for instance, the signal *SP* will be activated whenever the body of the robot is touched.

It is assumed that the system has a physical face with a repertoire of facial expressions that are commanded by a module that is not shown in Figure 2. When the robot is situated in front of a mirror, it will be able to perceive its own face visually. Initially it will not be able to connect the perceived face to any self-concept of its own; it will not recognize itself as the owner of the mirror image. However, this ability may arise from a series of steps as will be shown in the following.

Figure 3 describes the steps in the learning of mirror image self-recognition in the cognitive architecture of Figure 2. The various signals are depicted as the function of time and the whole process is divided into three temporal steps.

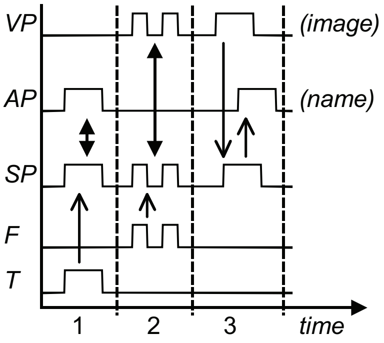

Figure 3. The learning of mirror image self-recognition

During step 1 the robot is touched and consequently cutaneous receptors will generate touch-signals *T*, which will cause a generalized somatosensory percept *SP*. At the same time the robot's name is announced. The corresponding auditory percept *AP* coincides with the somatosensory percept *SP* allowing the association of these with each other. In this way the name of the robot will be associated with the physical body of the robot; somatosensory percepts can evoke the name and the name can evoke generalized somatosensory percepts as if the body were touched in a general way. Consequently, the heard name of the robot would shift the robot's attention from the external world to the internal, on the robot's material self and whatever self-concepts were associated with it.

During step 2 the robot is placed in front of a mirror so that it can see its own image. It is assumed that the visual system detects the presence of some elementary and invariant features of the imaged scene and uses these as the actual visual percept instead of any raw pixel arrays. In this way the imaged objects do not have to look exactly the same each time when they appear.

Initially the robot will not be able to recognize the mirror image as its own reflection. If the robot now displays different facial expressions then these cause corresponding cutaneous signals *F*, which in turn will cause the generalized somatosensory percept *SP*. The facial expressions are reflected by the mirror and are perceived by the visual modality. The visual world contains also numerous non-relevant features and therefore any direct association of these with the somatosensory percept *SP* is not useful. Here the amodal features of the cutaneous percepts of facial expressions and the visual percepts of the face are used; if the timing of these features is used to control the association process ("attention control") then the correct association of the visual percepts of the face *VP* and the somatosensory percept *SP* will take place; the changed visual features (the face) will be associated with the changing somatosensory percept *SP*. This step grounds the visual mirror image self-recognition to the somatosensory basis of the robot's self-concept.

During step 3 the robot sees its mirror image *VP*, which is now able to evoke the robot's somatosensory self-concept *SP*. This in turn is able to evoke the name percept *AP* by the associative link that was acquired during step 1.

At this moment, instead of the mirror image a photograph of the robot could do as well. The robot would recognize itself in the photograph, that is, the image would evoke the name of the robot. However, the photographic image would be different from the mirror image, as it would not reproduce the motions of the robot. (This has implications to the case of robots with similar key features. In that case a robot may and should recognize the other robot as an "image" of itself, but not in the mirror image sense. Nevertheless, this might lead to the association of some of the robot's mental content with the other robot leading eventually to a theory of mind and social awareness. In this way a group of similar robots might spontaneously form a society of robots. Would this be another reason to build human-like robots; to allow robots to associate naturally with humans? More research would be needed along these lines.)

Mirror Self-Recognition and Consciousness

In the previous section a possible mechanism for mirror self-recognition is described. "Consciousness" is not mentioned there at all; consciousness is not required in the explanation and it is not claimed that this kind of mirror self-recognition would directly lead to any kind of consciousness, either. The author has argued that consciousness is not an entity or an agent that causes something; instead the folk-psychology hallmarks of consciousness arise from the cooperation of the various modalities in certain kinds of cognitive architectures. A similar idea has been also proposed by von der Malsburg (1997). Thus the subjective consciousness of a system is the way, in which the operation of the system appears to the system itself (Haikonen 2003, 2007b). Accordingly, mirror self-recognition is just an ability and process that facilitates certain kinds of skills and may provide contents to instantaneous consciousness if the system is otherwise able to support conscious processes. Especially, mirror image self-recognition would allow the generation of new visual percepts and inner models of bodily self, namely the visual model of one's face and other body parts that cannot be seen directly.

Thus, along these lines it should be easy to construct robots that have a somatosensorily grounded non-conscious self-reference with mirror self-recognition. This would be something like Gallagher's "minimal self" that would arise from the interaction between the robotic body and its environment (Gallagher 2000). According to the author's theory this non-conscious self-reference would turn into conscious one if it were to gain reportability and thus become introspected via unified system attention. This is effected by the cooperative cross-connections between the various modules and the feedback loops that return the evoked signals back to introspective percepts (more details Haikonen 2007b).

This mode of operation would liaise the instantaneous percepts with system needs and percepts from memory, which could be reported together; thus a "narrative self" (Gallagher 2000) could arise. The flow of reports ("thoughts" in general, not only verbal ones) would then be able to evoke further associations; imagery and possibilities for action and new focuses for attention.

Obviously the simplified cognitive architecture of Figure 2 is only able to associate simple self-related representations with sensory percepts and vice versa. Real consciousness and self-consciousness would include more complicated internal models (see e.g. Holland 2007); self-models, personal history as well as emotional evaluation and reactions. These self-models would include a "body image" that would allow knowledge about the position and motion of body parts and the imagination and prediction of the same. Another self-model would be a mental one, a "mental self-image". This model would integrate personal history with personal needs, goals, emotional values and styles of action; "who am I, what do I want, how do I react, am I good or bad, can I do this..."

The architecture of Figure 2 can be augmented towards these capacities, see Haikonen (2007b). The extended architecture would associate personal history and other self-related percepts with the somatosensory self-concept and all these could be cross-associated with each other. In this way an associative network of self-related entities and events would arise.

Is it possible to create machine consciousness by building machines that can recognize themselves in a mirror? Perhaps, but there is no shortcut here. The fundamental issues of qualia, the apparent immateriality of the mind and others must still be resolved in one way or another (Haikonen 2007a). The “Aleksander axioms”, a list of cognitive functions that are seen essential to consciousness (Aleksander and Dunmall 2003, Aleksander and Morton 2007) could also be a useful checklist.

Conclusions

Trivial mirror image self-detection in electronic systems is easy to achieve, but this function alone does not involve or cause any kind of self-awareness or consciousness. The author proposes that true self-awareness should involve a self-concept or a self-model that contains a “body image” and a “mental self-image” which are grounded to the material self via a somatosensory system. In cognitive robots an equivalent of the somatosensory system with cutaneous receptors should be realized and this system should be used to provide the grounding for self-related entities. This kind of self-concept grounding can be realized through the Haikonen cognitive architecture (and other functionally similar architectures). True mirror image self-recognition would evoke the robot’s self-model and facilitate the utilization of the correspondence of the mirror image and this self-model in possible motor actions. Self-consciousness would arise, when the self-percepts and self-models become reportable via unified system attention. This, in turn, calls for cooperation via associative cross-connections between the various modules of the system.

Acknowledgements

I wish to thank the head of the Nokia Research Center, Dr. Bob Iannucci and my superiors, Dr. Juha Lehikoinen and Dr. Kari Laurila for the possibility of working in this interesting field. I also thank Mr. Dylan Drummond for his most useful comments.

References

Aleksander, I., Dunmall, B. 2003, Axioms and Tests for the Presence of Minimal Consciousness in Agents. In Holland, O. ed., *Machine Consciousness*: 7–18. UK: Imprint Academic.

Aleksander, I., Morton, H. 2007. Why Axiomatic Models of Being Conscious? *Journal of Consciousness Studies* 14 (7): 15–27.

Damasio, A. R. 2000. *The Feeling of What Happens*. Vintage, UK

Fitzpatrick, P. Arsenio, A. 2004. Feel the beat: using cross-modal rhythm to integrate perception of objects, others, and self. Retrieved on 4.4.2007 from *http://groups.csail.mit.edu/lbr/mars/pubs/fitzpatrick04feel.pdf*

Fitzpatrick, P. Arsenio, A., Torres-Jara E. R. 2004. Reinforcing Robot Perception of Multimodal Events through Repetition and Redundancy and Repetition and Redundancy. Retrieved on 4.4.2007 from *http://people.csail.mit.edu/etorresj/PubDownload/FitzArseTorr05.pdf*

Gallagher, S. 2000. Philosophical conceptions of the self: implications for cognitive science. *Trends in Cognitive Science* 4 (1): 14–21.

Gallup G. G., Jr. 1970. Chimpanzees: Self-recognition. *Science* 167: 86–87

Gold, K., Scassellati, B. 2005. Learning about the Self and Others Through Contingency. Retrieved on 4.4.2007 from *http://gundam.cs.yale.edu/KgoldDevRob05.pdf*

Gold, K., Scassellati, B. 2006. Deictic Learning and Mirror Self-Identification. Retrieved on 23.3.2007 from *http://www.csl.sony.fr/epirob2006/papers/GoldScassellati.pdf*

Haikonen, P. O. 2003. *The Cognitive Approach to Conscious Machines.* UK: Imprint Academic.

Haikonen, P. O. 2007a. Essential Issues of Conscious Machines. *Journal of Consciousness Studies* 14 (7): 72–84.

Haikonen, P. O. 2007b. *Robot Brains: Circuits and Systems for Conscious Machines.* UK: Wiley & Sons.

Holland, O., Goodman, R. 2003. Robots with Internal Models: A Route to Machine Consciousness? In O. Holland (Ed.), *Machine Consciousness*: 77–109. UK: Imprint Academic.

Holland, O. 2007. A Strongly Embodied Approach to Machine Consciousness. *Journal of Consciousness Studies* 14 (7): 97–110.

Kawamura, K., Dodd, W., Ratanaswasd, P., Gutierrez, R. 2005. Development of a Robot with a Sense of Self. *The Proceedings of 6th IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA 2005*: 211–217

Lewis, M. Sullivan, M. W., Stanger, C., Weiss, M. 1989. Self-development and self-conscious emotions. *Child Development* 60 (1): 146–156

Malsburg, von der, C. 1997. The Coherence Definition of Consciousness. in Ito, M. Miyashita, Y. Rolls, E. T. eds, *Cognition, Computation and Consciousness,* UK: Oxford University Press

Michel, P., Gold, K., Scassellati, B. 2004. Motion-Based Robotic Self-Recognition. *The Proceedings of 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*: 2763–2768

Takeno, J., Inaba K., Suzuki T. 2005. Experiments and examination of mirror image cognition using a small robot. *The Proceedings of the 6th IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA 2005*: 493–498