




Robot in the Mirror: Toward an Embodied Computational Model of Mirror Self-Recognition

Matej Hoffmann¹ · Shengzhi Wang² · Vojtech Outrata¹ · Elisabet Alzueta³ · Pablo Lanillos⁴ 

Received: 1 June 2020 / Accepted: 22 December 2020 / Published online: 21 January 2021
© Gesellschaft für Informatik e.V. and Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

Self-recognition or self-awareness is a capacity attributed typically only to humans and few other species. The definitions of these concepts vary and little is known about the mechanisms behind them. However, there is a Turing test-like benchmark: the mirror self-recognition, which consists in covertly putting a mark on the face of the tested subject, placing her in front of a mirror, and observing the reactions. In this work, first, we provide a mechanistic decomposition, or process model, of what components are required to pass this test. Based on these, we provide suggestions for empirical research. In particular, in our view, the way the infants or animals reach for the mark should be studied in detail. Second, we develop a model to enable the humanoid robot Nao to pass the test. The core of our technical contribution is learning the appearance representation and visual novelty detection by means of learning the generative model of the face with deep auto-encoders and exploiting the prediction error. The mark is identified as a salient region on the face and reaching action is triggered, relying on a previously learned mapping to arm joint angles. The architecture is tested on two robots with completely different face.

Keywords Self-recognition · Robot · Mirror test · Novelty detection · Predictive brain · Generative models

1 Introduction

The “Turing test” of self-awareness or self-recognition was independently developed for chimpanzees [29] and infants [4] and consists in covertly putting a mark on the faces of

the subjects, placing them in front of a mirror, and observing their reactions. *Mirror self-recognition (MSR)* is often used to denote this test. The details of the mark placement, testing procedure, and assessment differ depending on the tradition [10]. In infants, a spot of rouge is covertly applied alongside the infant’s nose by the mother. Several behaviors may be counted as passing the test: ‘Touch spot of rouge’, ‘Turns head and observes nose’, ‘Labels self (verbal)’, or ‘Points to self’ [4]. In chimpanzees, a dye is applied to the unconscious animal and placed on *two* nonvisible locations (brow ridge and opposite ear). The assessment criteria are either ‘Combination of changing behaviors’ (decrease social responses, increase self-directed responses) or ‘Touches to mark’ [29].

The test has, in different variants, been used many times and in different species. Often it was interpreted in a binary fashion—specific species (humans, chimpanzees, orangutans, bottlenose dolphins, Asian elephants, Eurasian magpies) can pass the test and hence “possess self-awareness”, while other species do not. In humans, it is studied in a developmental perspective: infants pass the test at around the age of 20 months. However, de Waal [19] argues against such “Big Bang” theory of self-awareness and advocates a gradualist perspective instead.

M. H. and V. O. were supported by the Czech Science Foundation (GA ČR), Project nr. 17-15697Y. P. L. was partially supported by the H2020 project Selfception (nr. 741941).

✉ Pablo Lanillos
p.lanillos@donders.ru.nl
Matej Hoffmann
matej.hoffmann@fel.cvut.cz
Shengzhi Wang
shengzhi.wang@tum.de

- ¹ Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic
- ² Technical University of Munich, Munich, Germany
- ³ Center for Health Sciences, SRI International, Menlo Park, CA 94025, USA
- ⁴ Donders Institute for Brain, Cognition and Behaviour, Department of Artificial Intelligence, Radboud University, Nijmegen, The Netherlands

Despite 50 years of study of MSR, little is known about the mechanisms that bring about success in the test. A notable exception is Mitchell [46], proposing two theories. In this work, we follow up on this approach, introducing more detail, and attempt to unfold the MSR phenomenon into a block diagram, listing all the necessary prerequisites and modules. Furthermore, following the synthetic methodology (“understanding by building”) [35, 51] and the cognitive developmental robotics approach (e.g., [8, 14]), we will realize MSR on a humanoid robot, adding to the efforts to understand body representations and the self by developing embodied computational models thereof using robots (see [32, 36, 39, 42, 54] for surveys). Understanding MSR—still a relatively low-level milestone of the development of “self-knowledge” in humans [48]—will specifically generate insights into the sensorimotor, ecological [48], or minimal [27] self.

This article is structured as follows. Section 2 describes the mirror mark test, the possible mechanisms behind MSR, how humans recognize their own face, and summarizes previous works on self-recognition in robotics. Section 3 discusses in detail the mechanisms involved in MSR and presents a process model. In Sect. 4, we describe the methods needed for implementation on a humanoid robot—learning the appearance representation and detecting the mark in particular. Section 5 demonstrates quantitative and qualitative performance of the architecture components and the robot behavior under MSR on two versions of the robot Nao. Discussion is followed by Conclusion and future work.

2 Related Work

2.1 The Mirror Mark Test

The mirror mark test was independently invented for chimpanzees [29] and infants [4]. Bard et al. [10] provide an excellent overview of the details of the test in the different traditions. Comparative studies (e.g., [6, 19, 29]) have asked whether chimpanzees, orangutans, elephants, magpies, etc. as a species possess a self-concept; developmental studies (e.g., [4, 5]) have been concerned with individual differences and developmental milestones. In both traditions, the mirror mark test is a gold standard, an objective assessment, appropriate for nonverbal or preverbal organisms, relying on objective target behavior: reference to the mark on the face, after discovering the mark by looking in the mirror [10]. However, there are differences in the mark application (infants: spot of rouge applied covertly by mother and placed in a single location alongside nose; chimpanzees: alcohol-soluble dye applied to unconscious animal in two non-visible locations) and testing procedures (e.g., infants: mother scaffolds infant’s response). The biggest difference among

Table 1 Assessment criteria for mirror self-recognition from [10]

Amsterdam (1972)	Gallup (1970)
Recognition of mirror image:	Combination of changing behaviors:
-Touch spot of rouge	-Decrease social responses
-Turns head and observes nose	-Increase self-directed responses
-Labels self (verbal)	Touches to mark:
-Points to self	-Confirms self-directed touches
	-More when mirror present than absent

experiments is probably the interpretation of the responses and judging whether the test has been passed (reference to the mark). According to Mitchell [46], “the standard evidence of ‘mirror-self-recognition’ is an organism’s responding, when placed before a mirror, to a mark on its forehead or other area of the body which is not discernible without the use of the mirror” [4, 29]. However, the detailed assessment criteria differ, as summarized in Table 1 (from [10]). For the purposes of this article, in which we seek explanations for the most low-level, sensorimotor, aspects of MSR, it is the “touch spot of rouge/mark” that will be our focus.

2.2 Mechanisms of Mirror Self-Recognition

What does success in MSR entail? Mitchell [46] proposed two theories, or mental models, which are schematically illustrated in Fig. 1: the “inductive” and the “deductive” theory. The “inductive theory” (in agreement with the observations of [31]) presumes that subjects that are (1) capable of visual-kinesthetic matching and that (2) understand mirror correspondence are likely to pass the mirror test. The “deductive theory” [46] should be stronger: if (1) full understanding of object permanence, (2) understanding mirror correspondence, and (3) objectifying body parts is in place, it constitutes a necessary and sufficient condition for MSR. For evidence supporting the theories, the reader is referred to [46].

Let us first look at the inductive theory in detail. Kines-thesia means movement sense but is sometimes equated with proprioception: the subject is aware of where its body is in space based on somatosensory afference. Visual-kinesthetic matching means that the subject can map this information onto a visual image: how such a body configuration would look like. This is necessary for imitation, which is why imitation capabilities are monitored in relation to the likelihood of passing the mirror test. Specifically for MSR, kinesthetic-visual “self-matching” is needed. Mitchell [46] is not completely clear whether this capacity is primarily spatial (comparing static kinesthetic and visual body configuration images) or temporal (both “images” moving in synchrony),

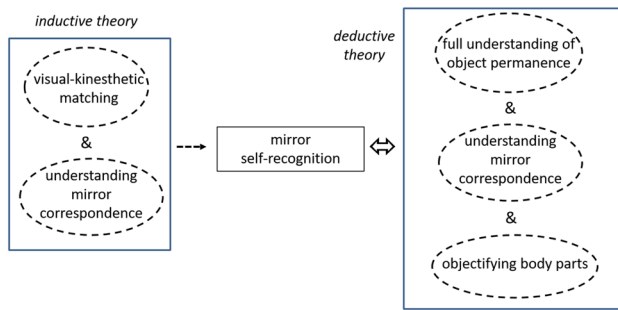


Fig. 1 Inductive and deductive theory of mirror self-recognition according to Mitchell [46]. The two theories are enclosed in rectangular blocks, with individual capacities enabling MSR in dashed ellipsoids. The arrow from inductive theory to MSR is dashed as the capacities are likely but not sufficient to enable MSR. The deductive theory provides necessary and sufficient conditions and hence the arrow denoting equivalence

but it seems that emphasis is on the former. Adding understanding mirror correspondence—that mirrors reflect accurate and contingent images of objects in front of them—allows the subject to add the mapping needed between the visual body image she constructed and the mirror reflection.

The deductive theory lists three conditions that are necessary and sufficient for MSR. The first prerequisite is full understanding of object permanence, corresponding to Piagetian stage 6 of this capacity [52], which “presupposes that an organism has memory and mental representation, recognizes similarity between similar objects, recognizes that its body is a continuous object localized and extended in space (and therefore *represents its body as such* in this way), ..., and has some primitive deductive abilities.” [46] The second prerequisite, understanding mirror correspondence, has been discussed above. By objectifying body parts, the third capacity, “is meant both (1) that the organism recognizes the similarity between any particular body-part of-its own and the same body-part of another (a recognition presupposed by understanding object permanence), and (2) that the organism recognizes a body-part, and recognizes it as part of the body, even when the body-part is decontextualized—that is, separated from the body.” [46] Then, according to Mitchell, “the organism perceives *x* (its hand) as an object which is distinct yet continuous with *y* (its body), and knows that mirrors reflect objects accurately and contingently; if *x* is an object distinct yet continuous with *y*, and if mirrors reflect objects accurately and contingently, then if a mirror reflects *x*, it must simultaneously reflect *y*; the organism knows that the mirror reflects *x*; therefore the organism knows that the mirror reflects *y* and thus recognizes the mirror-image as an image of its body.” [46]

Passing the test without “cheating” assumes, first, that the subject identifies herself in the mirror (“it’s me in the mirror” box). That is, if the subject thinks it is her conspecific in

the mirror with a mark on the forehead and then goes on to explore also her forehead, it should not count as passing the test and should be controlled for. Gallup [28] postulates “an essential cognitive capacity for processing mirrored information about the self”. Animals that possess this capacity or that “are self-aware” can succeed. Alas, such an explanation does not bring us closer to understanding the mechanisms. Anderson [5] assumes that recognition of one’s body-image in a mirror results from “a mental representation of self onto which ... perception of the [mirror] reflection is mapped” (cited from [46]).¹

However, a “self-image” is not the only way of getting at “it’s me in the mirror”. “Temporal contingency (image moves as I move)” may presumably be more effective—bypassing the problem of visual matching dependent on image translations, rotations, size, clothes in case of infants etc. This is an instance of the general question of body ownership vs. agency [62]: which of them, or perhaps their combination, is relevant to pass the “it’s me in the mirror” in the MSR context? Bigelow [11] provides evidence for the temporal contingency cue in a study where movement was used as a cue to self-recognition in children by presenting them with movies of their own or other infants’ photographs in or out of sync with their movements. Thus, an image of how one’s whole body looks from the outside may not be necessary: the subject may do away with identifying it is her through temporal contingency and then an image of her face may suffice. However, in practice, temporal and visual contingency are intertwined during the MSR test.

To succeed in the test, the subject needs to display one of the behavioral responses listed in Table 1. To display a response, the subject needs “motivation”, which has also generated some controversy in the field: the responses should be spontaneous and engineering them in species that do not normally pass the test through reinforcement—monkeys in particular—has been criticised [6, 19]. The response we focus on is “touch spot of rouge”/“touch to the mark”. There are several ways of reaching for the mark. One key distinction is whether the reaching is “feedforward” or “feedback”. In the study of reaching development, the “visually guided reaching” hypothesis—infants need to look at their hands and the object alternately in order to progressively steer the hand closer to the object location—has been replaced by “visually-elicited” reaching, whereby the infant looks at the target and continues to do so during the reaching (Corbetta et al. [18] provide an overview and additional evidence). If the latter strategy was employed to reach for the

¹ Mitchell [46] discusses the “chicken-and-egg problem” of acquiring this self-image—prior recognition in the mirror may be necessary to learn it—and concludes that there are “three possibilities: (1) a visually based, incomplete self-image of the part of the organism it can see, (2) a non-visual self-image, (3) or a mixture of these images.”

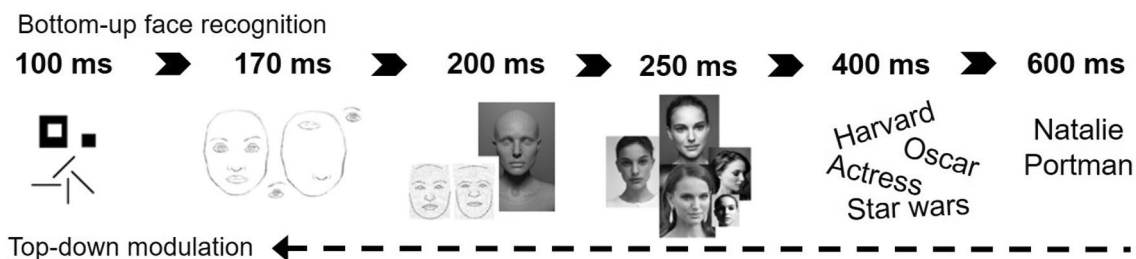


Fig. 2 Human face self-recognition process

mark, first, some form of “remapping of stimulus location in the mirror to an ego-centric frame” is necessary. Heed et al. [34] review the process of “tactile remapping”: how a stimulus on skin may be transformed such that it can be acted upon (looked at, reached for, ...). However, how that is achieved in the case of the mirror remains an open question. Some “understanding of mirror correspondence” is probably recruited. However, it seems very unlikely that such understanding will be so mature that the mark could be localized in space with respect to the body frame. Instead, it seems more likely that the mark location is identified with respect to some landmarks on the face (nose, mouth, etc.) encoded in lower hierarchical levels of visual face recognition [22]. Then, the rest of the localization process may be similar to tactile remapping.

The role of the tactile modality remains to be further explored. The experiments of Chang et al. [15] with macaques show that making the association between the visual stimulus in the mirror and tactile sensation (using high-power laser) facilitates later success in the mirror test. Perhaps, this association is needed to “bootstrap” the remapping process. This is hypothesized also by de Wall [19]: “it is as if these animals need multimodal stimulation to get there” (also called “felt mark test” in [20]).

An alternative approach to MSR would be to consider the brain as a Bayesian machine that encodes sensory information in a probabilistic manner. In this scheme, self-recognition is achieved by differentiating our body from other entities because they are probabilistically less likely to generate the observed sensory inputs [7]. This inevitably requires the ability of encoding priors (by learning) and the availability of generative processes that predict the effects of our body in the world [24, 43]. For instance, within this approach, visual congruency may be computed through minimizing the prediction error between the expected appearance and the current visual input at each hierarchical layer—a predictive coding approach [55]. Under the Free-energy principle [7, 24], self-recognition is related to low surprise in the sensory input. The brain is able to correctly predict and track the body effects in the world.

The behavioral response and its awareness in the MSR is, however, a controversial aspect. Friston [24, 25] proposed that the action minimizes the prediction error derived from proprioceptive expectations by means of the classical motor reflex pathway. In [65], human participants exerted involuntary movements during a sensory conflict experiment to reduce the sensory prediction error. Hence, we hypothesize that salient regions in the face (e.g., mark) will produce a goal driven response to reduce this visual “error”, and therefore, we would expect “visually guided reaching” as described above. It is still debatable that the behavioral response produced in this scheme will encompass self-awareness as studied in primates. This will require some level of body-ownership and agency using the mirror reflection [64].

2.3 Face and Self-Face Recognition in the Brain

According to the cognitive and functional model of facial recognition in humans [13, 37], recognition occurs through a hierarchical process, as depicted in Fig. 2. Initially, facial low-level visual information is analyzed. Then, facial elements (eyes, nose, mouth) are detected [22] and the spatial relationship between them are integrated forming a layout [44]. Once the physical characteristics of the face have been coded by the visual system, the resulting representation must be compared with the known faces representations stored in long-term memory—at the Facial Recognition Units (FRUs) [58]. Only when there is high correspondence between the computed and the perceived representation, there is access to the semantic and episodic information related to the person (relationship, profession, etc.), and finally to his/her name [58]. This last stage happens at the so-called Person Identity Nodes (PINs), which can be activated by different sensory pathways (e.g., auditory by someone’s voice).

This general face recognition slightly differs when recognizing our own face. Studies that investigate the temporality of self-face recognition show that self-face differs from general face recognition already at an early stage in visual processing (around 200 ms after the stimulus onset) [3]. At this stage, self-face processing is characterized by a reduced

need for attentional resources. This bottom-up attention mechanism facilitates the activation of self-face representation on memory (FRUs) and therefore self-recognition. Surprisingly, once self-face has been recognized, a top-down attentional mechanism comes into play allocating cognitive resources on the brain face-related areas to keep self-face representation in active state [2].

Neuroimaging studies also evidence the interplay between bottom-up and top-down attentional control brain areas during self-processing [60]. The activation of a specific Self-Attention Network supports the theoretical view that attention is the cognitive process more distinctive during self-face recognition. In the case of the mirror test, this increased attentional mechanism would strengthen the relevant signals such as novel visual cues on the face (e.g., mark) as well as boost the access to memory and produce the feeling of awareness.

2.4 Robot Self-Recognition

Several works have addressed self-recognition in robots (works until 2010 reviewed in [36] and revisited in [42] from the enactive point of view). First, we describe works on body self-recognition and second, we summarize works that specifically studied robots in front of a mirror.

Two principally different strategies were employed for machine self-recognition. According to the first, the body is the invariant: what is always there. The research of Yoshikawa and colleagues (e.g., [63]) capitalizes on this property, acquiring a model of “how my body looks like”. In [21, 40], the robot learns the appearance of its hand and arm using deep learning techniques. The second strategy takes a largely opposite approach: my body is what moves, and, importantly, what I can control. Fitzpatrick and Metta [23] exploit the correlation between the visual input (optic flow) and the motor signal; Natale et al. [47] improve the robustness of this procedure by using periodic hand motions. Then, the robot’s hand could be segmented by selecting among the pixels that moved periodically only those whose period matched that of the wrist joints.

Bayesian and connectionist approaches have been proposed to capture this sensorimotor correlation for self-recognition. Tani, in [61], presents self-recognition from the dynamical systems perspective using artificial neural networks. Gold and Scassellati [30] employ probabilistic reasoning about possible causes of the movement, calculating the likelihoods of dynamic Bayesian models. A similar approach was proposed in [41, 50], where the notion of body control was extended to sensorimotor contingencies: “this is my arm not only because I am sending the command to move it but also because I sense the consequences of moving it”. All these exploited the spatio-temporal contingency, related to the sense of agency. Pitti

et al. [53] studied temporal contingency perception and agency measure using spiking neural networks. Gain-field networks were employed to simultaneously learn reaching and body “self-perception” in [1].

Specifically, mirror self-recognition has also been studied in robots. Steels and Spranger [59] explored this situation from the perspective of language development. A Nao robot was engineered to pass the mirror test using logical reasoning in [12]. Hart [33] employed state-of-the-art techniques in computer vision, epipolar geometry, and kinematics. Fuke et al. [26] proposed a model in which nonvisible body parts—the robot’s face—can be incorporated into the body representation. This was done via learning a Jacobian from the motor (joint) space to the visual space. A neural network with Hebbian learning between the visual, tactile, and proprioceptive spaces was used. Integrating the velocities, position in visual space can be estimated for nonvisible parts as well. Then, while the robot was touching its face with the arm, the position in the visual modality could be estimated and matched with the touch modality, learning a cross-modal map.

Finally, Lanillos et al. [43] recently proposed an active inference (i.e. free-energy principle) approach to MSR where the robot learned to predict the sensory consequences of its actions using neural networks in front of the mirror and achieved self-recognition by means of evidence (absence of prediction error) accumulation.

In most of the works cited above, self-recognition in general or MSR in particular was largely engineered. In this work, we present a process model that is more tightly grounded in the psychological literature on MSR. Furthermore, we present an embodied computational model on a humanoid robot, in which the novelty detection is currently our main contribution.

3 Process Model of Mirror Self-Recognition

Mitchell’s theories ([46] and Sect. 2.2) suggest possible modules or components that may be needed for MSR. However, they are still quite abstract high-level capacities and their specific role in the *process of passing the mirror test* is unclear. Therefore, we propose instead a *process model* (Fig. 3) of going through MSR. We do employ “visual-kinesthetic matching” and “understanding mirror correspondence” blocks in our model, while we leave “fully understanding object permanence” and “objectifying body parts” aside—these are very complex capacities that would be far from trivial to implement. Instead, we hope that their explicit instantiation is not necessary for MSR. Possibly, behavior that can be interpreted as such in this context may emerge in our model. Our proposed mechanistic account is shown in Fig. 3.

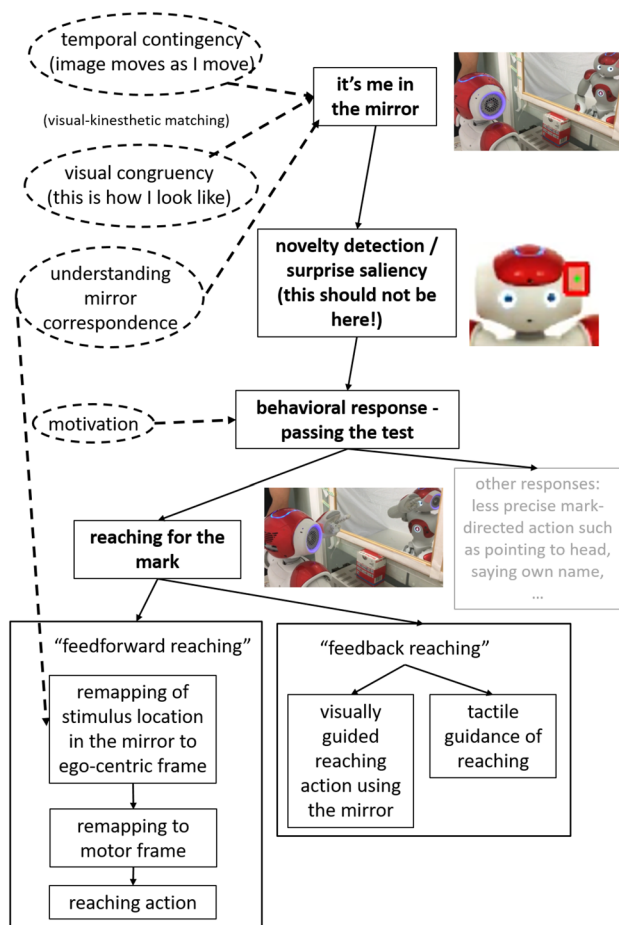


Fig. 3 Process model of mirror self-recognition. The main pipeline is illustrated using blocks with text in bold. From top to bottom: self-recognition in the mirror, identifying the mark, reaching movement. The control of the reaching movement can be feedforward (left) or feedback (right). Circles with dashed lines illustrate “prerequisites” for individual blocks

3.1 It’s Me in the Mirror

The first “block” in the process model of MSR is “it’s me in the mirror”, which likely arises from some visual-kinesthetic matching as discussed in Sect. 2.2, and takes the form of “visual congruency” (“this is how I look like”) stressed by [46], or “temporal contingency”, or their combination. Additionally, “understanding mirror correspondence” likely also contributes to the “it’s me” test. It seems plausible to think that it would be relatively more important for the “visual congruency” cue, whereby the image of the self needs to be matched against its specular reflection.

3.2 Novelty Detection/Surprise Saliency

The “novelty detection/surprise saliency” is the module responsible for recognizing one’s own face (Sect. 2.3) and

detecting the mark as an object that does not belong there. It is this part that our computational model on the robot will specifically address (Sect. 4)—constituting the main technical contribution in this work.

3.3 Reaching for the Mark

To succeed in the test, the subject needs to display one of the behavioral responses listed in Table 1. Our focus will be the “reaching for the mark” and we will leave the “other responses” aside, as these do not easily lend themselves to a mechanistic decomposition. To display a response, the subject needs “motivation”. However, the responses should be spontaneous and not engineered through external rewards. From the perspective of our process model, the problem with such reinforcement is that the subject may learn to pass the test while side-stepping the “it’s me in the mirror” and “novelty detection” blocks.

How the “reaching for the mark” is performed remains an important open question. The subjects passing MSR should be capable of “feedforward reaching”. If this strategy is employed to reach for the mark, first, some form of “remapping of stimulus location in the mirror to ego-centric frame” is necessary. It is not clear how this is done in this case. “Understanding mirror correspondence” will facilitate the localization. However, it seems unlikely that such understanding will be so mature that the mark location could be remapped into, say, the body frame through a combination of stereo vision and mirror projection. Instead, it seems more likely that the mark location is identified with respect to some landmarks on the face (nose, mouth, etc.). Then, the rest of the localization process may be similar to tactile remapping. Next, the target location may be transformed to “motor coordinates” and finally executed.

Alternatively, it could be that in this unusual situation, the subjects would employ a “visually guided reaching action using a mirror”. Furthermore, in case the initial reach is not accurate, “tactile guidance of reaching” is also a possibility. All these options can in principle be realized in a robot. However, more information from experiments with infants and animals is needed to provide the right constraints for the model.

4 Mirror Recognition in a Robot

In this section, we present the architecture that we implemented on a humanoid robot—Fig. 4. Compared to Fig. 3, it is in many ways simplified. The novelty detection/surprise saliency was modeled in detail, inspired by the predictive coding hypothesis. In order to simplify the system, we will assume that the robot is able to identify

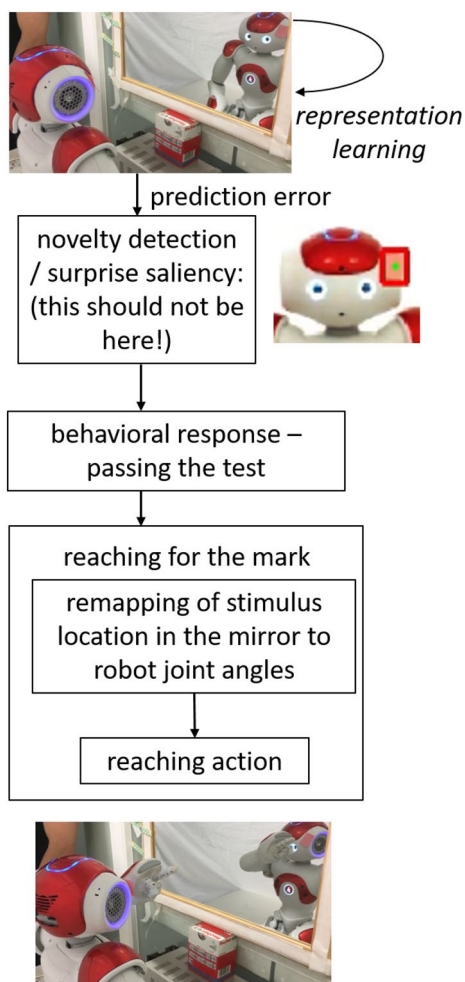
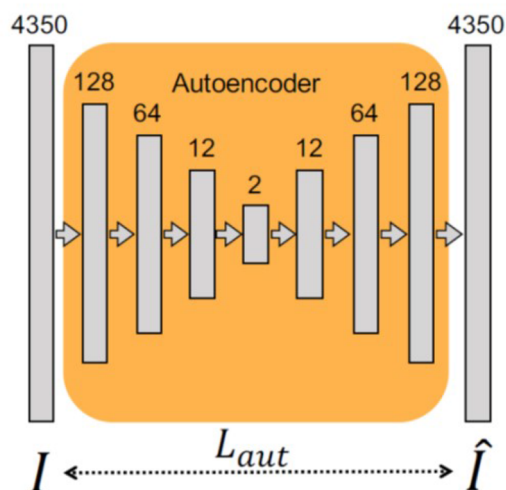


Fig. 4 Schematics of mirror self-recognition implemented in this work on the robot

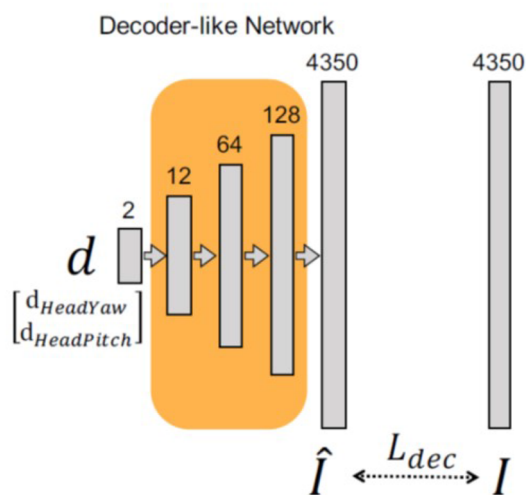
itself in the mirror, triggering the top-down modulation and attentional capture. That is, the high-level hierarchies of the face recognition process described in Fig. 2 are not addressed. Besides, the reaching for the mark has been grossly simplified at the moment.

4.1 Learning the Appearance Representation

We assume that the robot has some kind of self-distinction abilities using non-appearance cues [41, 43]. Therefore, it can learn the face representation in front of the mirror through semi-supervised learning. This allows the robot to imagine or predict its face appearance in the visual space. We studied two different network architectures described below and depicted in Fig. 5.



(a) Autoencoder



(b) Decoder

Fig. 5 Generative model learning. Two convolutional neural networks were tested: **a** autoencoder and **b** decoder with joint angles as the input. The number of nodes for each layer used is detailed

4.1.1 Self-Supervised Autoencoder

The first architecture, depicted in Fig. 5a is known as an autoencoder [9]. We used this artificial neural network (ANN) to learn the high level visual features of the face, analogously to the visual face recognition in humans. The

network also learns the generative process or “visual-kinesthetic” forward model predictor. Given the input image I and the predicted image \hat{I} , the network was trained using the MSE reconstruction loss: $L = \frac{1}{N} \sum (I - \hat{I})^2$, where N is the number of pixels in the image. We used tanh as the activation function in each layer except the output layer, where the Sigmoid function was selected to obtain the desired pixel output in the range of (0, 1).

4.1.2 Supervised Decoder-Like Neural Network

Inspired by the decoder part of the autoencoder architecture, we built a decoder-like neural network, as shown in Fig. 5b, where the latent space was substituted by the joint encoders of the robot head. Similar architectures have been used for learning the visual forward model [40, 57]. We denote $\mathbf{d} = [d_{HeadYaw}, d_{HeadPitch}]$ the head motor state as input vector to the network. Analogous to the autoencoder approach, tanh was the activation function in each internal layer and Sigmoid as the output layer. The predicted image \hat{I} is compared with the original image I corresponding to the input motor state \mathbf{d} . We also used the MSE between \hat{I} and I as training loss. Here, the motor state generates the corresponding image.

4.2 Visual Novelty Detection Using Generative Models and the Prediction Error

Once the appearance representation is learned, the robot can directly use the generative model to discover novel visual events, such as a colored mark placed in the face. Surprising events will have high prediction error. Therefore, we computed the image saliency by subtracting the predicted visual input and the current observation ($I - \hat{I}$). In practice, once the generative model is trained, the robot can predict its visual appearance depending on the head angles and compute the visual prediction error. However, simply computing this difference would lead to inaccurate distinction between highly variable regions (e.g. the eyes, mouth, light reflections, etc.) from real novel visual events. We need to take into account the variance associated with each pixel information. Hence, we computed the distribution that encodes mean prediction error for each pixel μ and its variance σ^2 as follows:

$$\mu = \frac{1}{N_i} \sum_{i=1}^{N_i} |I_i - \hat{I}_i| \quad (1)$$

$$\sigma^2 = \frac{1}{N_i - 1} \sum_{i=1}^{N_i} (|I_i - \hat{I}_i| - \mu)^2 \quad (2)$$

where N_i denotes the number of collected images and \hat{I}_i is the i th predicted face generated by the network. Finally, the saliency map I_s was computed by means of the Mahalanobis distance (D_M):

$$I_s = D_M(I, \hat{I}) = (|I - \hat{I}| - \mu) / \sigma^2 \quad (3)$$

where all operations are pixel-wise, I, \hat{I} are current specular image and the predicted image, respectively. Note that we are assuming that there is no correlation within pixels and thus, the covariance matrix Σ is diagonal and it is defined by the σ^2 vector.

4.3 Reaching for the Mark

In the current model, the reaching behavior has been greatly simplified and engineered. It could be said the our solution corresponds to the “feedforward reaching” strategy of Fig. 3; however, no remapping from the image in the mirror to an egocentric reference frame is performed. Instead, the robot head and arm were manually driven into configurations in which they reach for the mark at different locations on the face. The mapping from the detected novelty region in the specular image to the robot joint angles is directly learned (Fig. 4). To this end, we used a feedforward neural network to map the centroid position of the mark in the visual space $c = (i, j)$ to joint states q (e.g., head yaw, shoulder pitch). The ANN had three layers: input layer with 2 neurons for the mark location in the mirror, one hidden layer (10 neurons) and output layer with the number of neurons matching the number of joint states. We used the Sigmoid as the activation function in each layer.

Once this mapping, $(i, j) \rightarrow \mathbf{q}$, is learned, the points in the image frame in pixels have a direct translation to the joint angles space. The target joint angles are eventually sent to the joint position control of the Nao robot and the movement is achieved via local PID motor controllers.

5 Experiments and Results

Here we describe the experimental setup and the implementation details of the model deployed. We quantitatively evaluated the face representation learning and novelty detection system and qualitatively the robot behavior with two different Nao robots. A supplementary video showing the system in action can be found here: <https://youtu.be/lbdAyOPkIIM>.

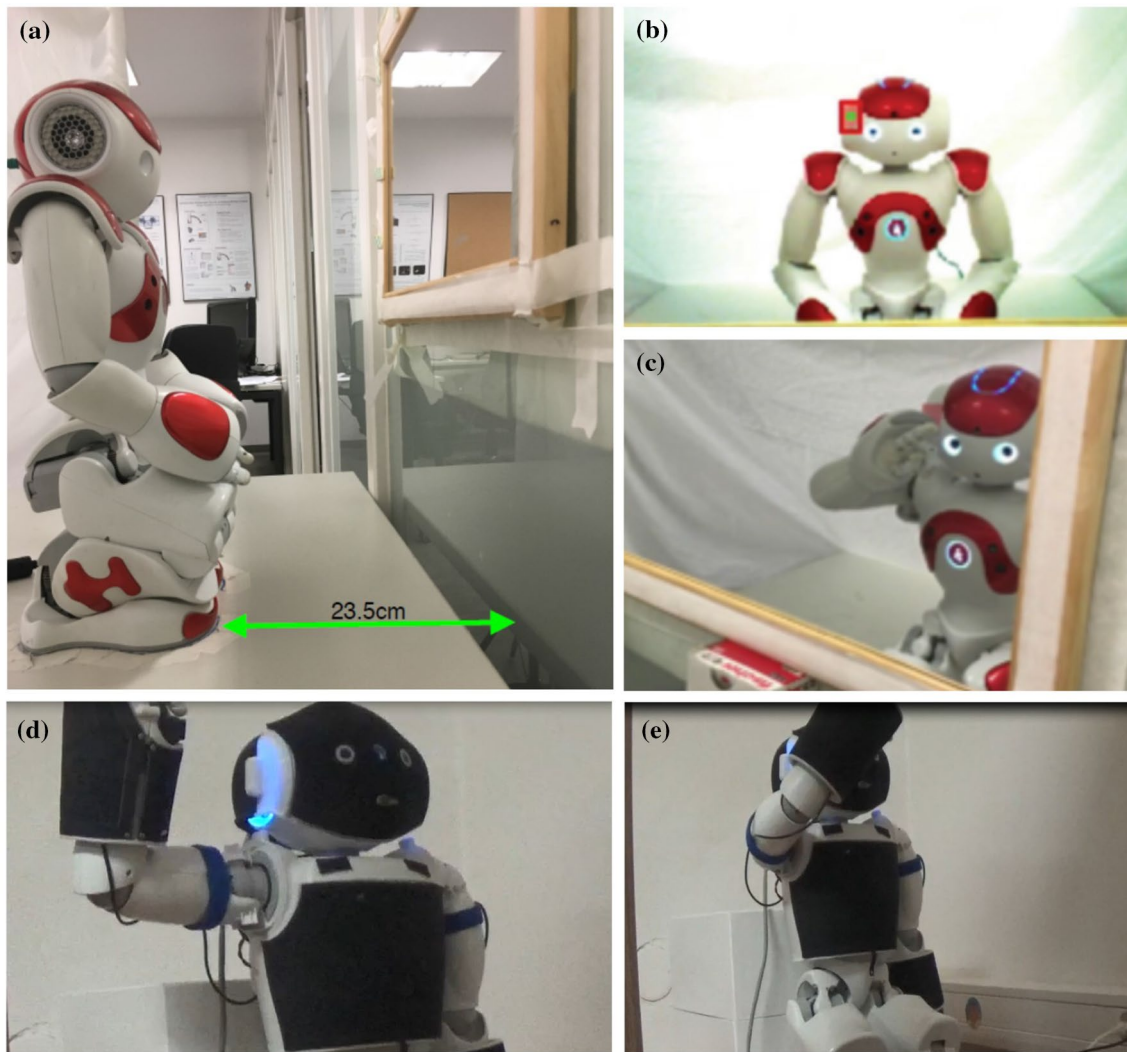


Fig. 6 Experimental setup. **a** A Nao robot was placed in front of the mirror for all trials. **b** Mark detection using the generative model prediction error approach. **c** Behavioral response towards the mark. **d**

Mirror reflection of the second Nao with electronic skin on the face. **e** Reaching behavior in the second Nao

5.1 Experimental Setup

The experimental setup is shown in Fig. 6. We tested the approach in two Nao robots with different appearance in both simulation and using the real platforms. We placed the robot facing a mirror at a fixed distance, leaving the face and the torso visible for both training and test trials.

5.2 Training and Evaluation

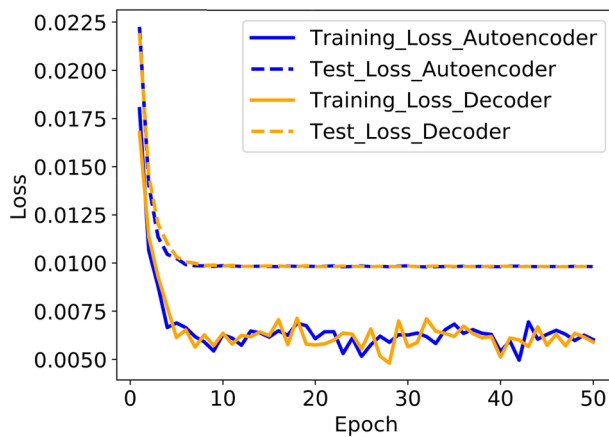
5.2.1 Learning the Face Representation

To evaluate the performance of the representation learning, we collected 1300 mirror reflection images in the Gazebo simulator and transformed into grayscale. Nao camera resolution is 1280×960 . During training, head yaw and pitch

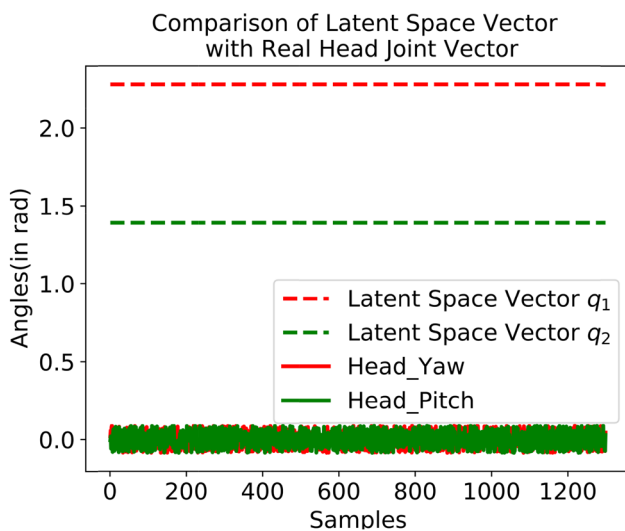
were randomly sampled in a range of -5° to 5° . The head region of the robot was cropped out by using the *OpenCV* function `matchTemplate` using an example robot head as a template.

We randomly selected 80% of the collected images as our training set, while the rest (20%) were assigned as test set. For the test dataset, we modified the images by synthetically adding a mark (e.g., a rectangle of 14×14 pixels) with random color (from a set of predefined colors) placed in a random position in each test image. The selection of the mark location and color was uniformly distributed.

We used the ADAM optimizer [38] to train both ANNs described in Sect. 4.1 and all hyperparameters were fixed equally for both models. Specifically, the number of training epochs was set to 50. A random mini-batch with the size of $N = 128$ was used to train the models for one iteration in



(a) Training and test loss



(b) Autoencoder latent representation

Fig. 7 Learning face representation—training and testing. **a** Autoencoder and decoder-like model loss on training and test dataset during the training process. **b** Autoencoder latent space vs. head joint state vectors. The encoded representation did not match the joint states

every epoch. The initial learning rate was set to 0.005, and it decreased after every 10 epochs by a decay factor 0.5.

Figure 7a shows the training loss curves for both training and test sets during the training process. Both models achieved similar reconstruction accuracy, converging to a MSE error of 0.01 on the test set. The autoencoder latent representation learned did not match the head orientation of the robot (i.e. head joint motor states $d_{HeadYaw}$ and $d_{HeadPitch}$), as shown in Fig. 7b.

5.2.2 Novelty Detection

After we trained the models, we compared the performance of the two different ANN architectures by evaluating the

accuracy for detecting the mark in the face. First, the saliency image I_s was binarized by selecting the pixels where the D_M (Eq. 3) was greater than 1.8% of the maximal pixel value in greyscale. Afterwards, areas which contained at least 30 consecutive saliency pixels were taken as relevant regions. The most salient area (in a winner-take-all manner) was selected as the output (i.e. mark) from the algorithm for evaluation purposes.

To evaluate the novelty detection performance, we defined the precision measure metric [41]: intersected area divided by the sum of intersected and detected area ($a_i/(a_i + a_d)$)—see Figure 8a. Values close to one mean that both regions overlap.

Figure 8b shows the accuracy comparison for the two generative models. The measure was calculated at every epoch by averaging its value over all the test set during the training process. Both architectures obtained similar results and converged to a steady state in less than 20 epochs. Results indicate that the prediction error variance was critical and more important than the prediction error in order to properly segment the mark from the face. In particular, when setting the prediction error variance to 1, many regions of the face become salient and the mark was not properly segmented returning low values of our metric.

Figure 9 shows five examples of the saliency I_s and the mark detection I_d computation using different mark colors and shapes. Although the highest saliency region corresponded to the mark, we can observe other parts of the face that have relevance, such as the face boundaries. Eyes had a high variance in the prediction error, affecting the mark detection.

5.3 Online Mirror Test on Real Robots

We finally tested our architecture on the real Nao robots. On the “standard”, red, robot (Fig. 6a–c), we first trained the face representation network with 800 real specular images. Second, we trained the visuo-proprioceptive mapping network (Sect. 4.3) by generating a database of visual mark locations and corresponding reaching joint states by manually moving the robot arm to reach close to the mark. To this end, we selected 5 joints² (*Head yaw, Head pitch, Left shoulder pitch, Left shoulder roll and Left elbow roll*) as the output of the ANN and we manually moved the robot arm and head to reach the mark in the specular image. The reaching space of the robot is constrained to the face in the training phase, thus, placing the mark outside the face could result in interpolated or undesired reaching behaviors. ADAM optimizer was used for the training; the learning rate was set to 0.01 and the total training iterations were 10^4 .

² Elbow pitch and wrist rotation were fixed.

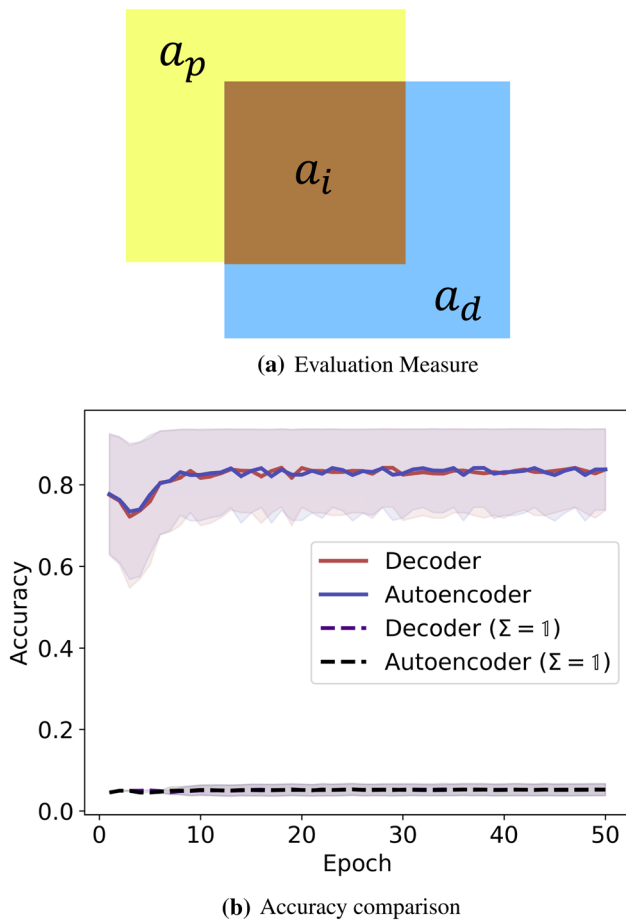


Fig. 8 Evaluation of the novelty detector. **a** Definition of the overlapping areas of the visual mark and the detected salient region: a_p denotes the area of the mark minus the intersected area a_i ; a_d denotes the detected area except the intersected area (i.e., false positives). **b** Accuracy comparison between the two ANN architectures with and without prediction error weighted by the variance ($\Sigma = \text{diag}(1)$)

After training, post-it notes of different colors and sizes were used as marks and their position was changed after every reaching response. Figure 10 shows examples from three trials. For these tests, we only used the decoder-like model as the generative process to predict the visual input.

To verify the robustness of our approach, we implemented the same architecture in a second robot—with a different setup and mirror and using a robot with very different visual appearance. Training had to be repeated. Figure 11 shows three executions with the other Nao robot. The columns show the camera registered image, the detected mark, and the corresponding reaching behavior respectively.

6 Discussion

The nature of our model is quite different than that of Mitchell [46]. The theories that Mitchell puts forth correlate success in MSR with other capacities: visual-kinesthetic matching, understanding mirror correspondence, object permanence, or objectifying body parts. However, testing each of these is, first, a challenge in itself. Second, such evidence still falls short of explaining the mechanisms of MSR. Instead, our modeling targets the process of undergoing MSR by developing an embodied computational model on a robot in front of a mirror. We do not treat MSR as a binary test—where one can succeed or fail and where success can be achieved (or engineered) in many ways—but as a process in which the behavioral manifestations of getting through the test can help us uncover the putative mechanisms. In this work, we model the mark detection on the face in some detail. Using novelty detection in the image of one’s face as the behavioral trigger is an assumption we are making, inspired by the attention system studies in humans [17]. The relation of our model to active inference [7, 24, 43] lies in its

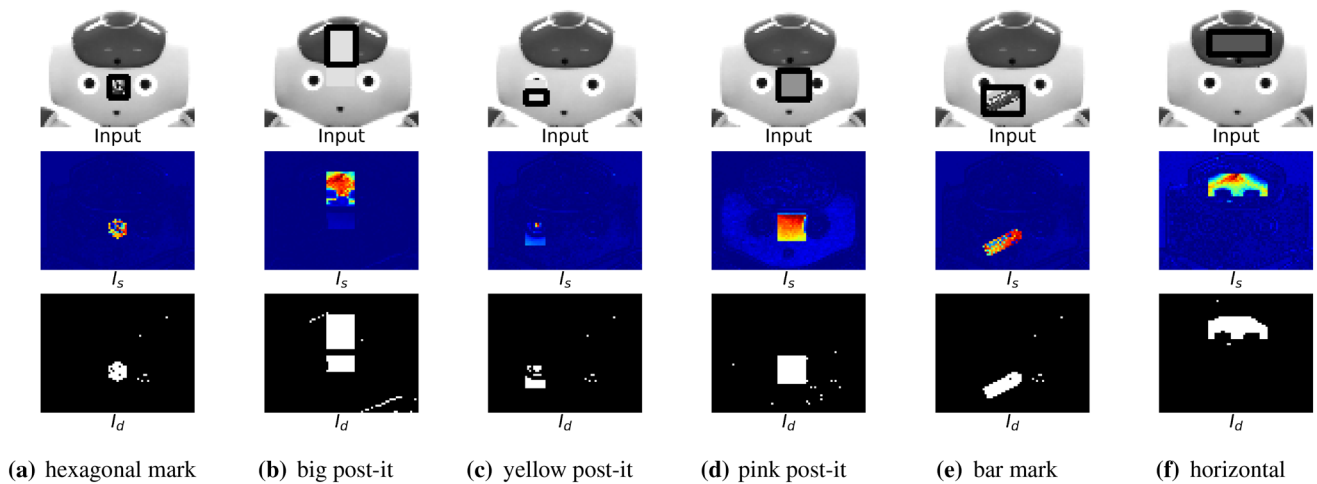


Fig. 9 Saliency and mark detection examples. The input image is the automatically cropped face; I_s is the computed saliency; I_d is the binarized saliency, i.e., mark detection output

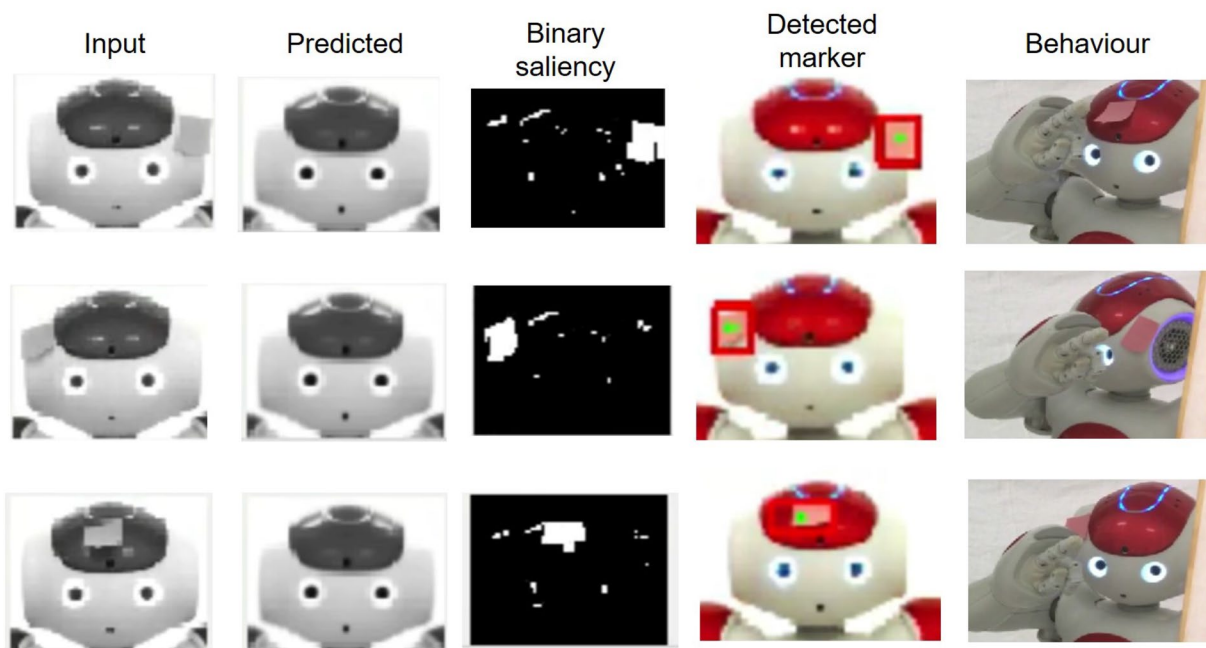
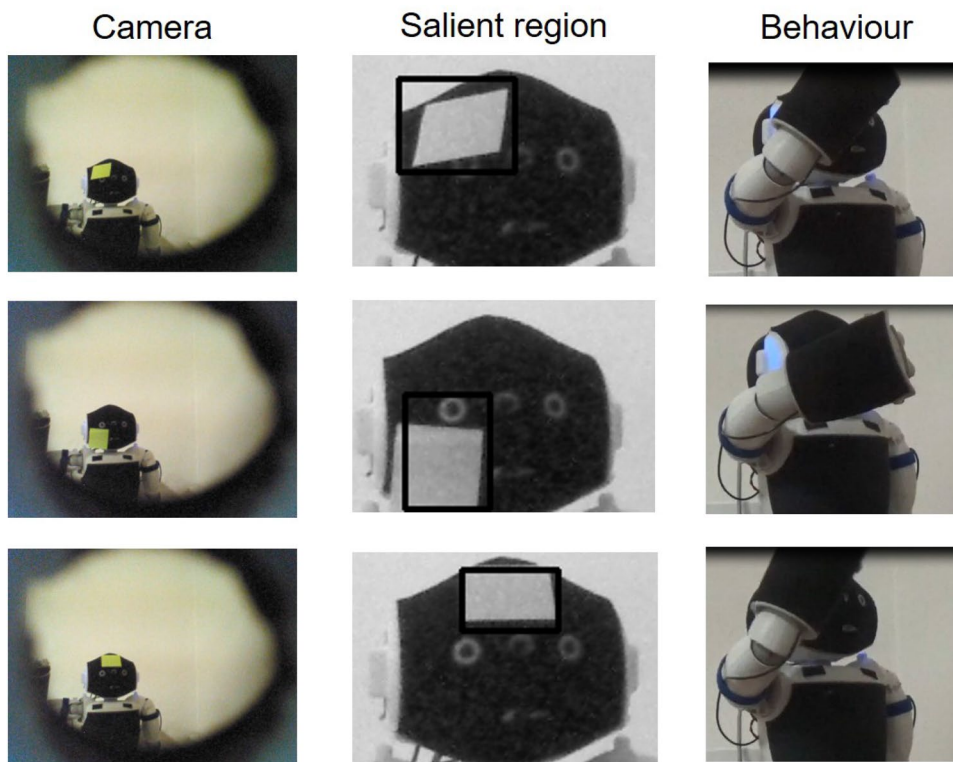


Fig. 10 Examples of passing the mirror test with the Nao robot. The input corresponds to the input image to the ANN and the output is defined by the reaching behavior

Fig. 11 Examples of passing the mirror test with a second Nao robot with custom covers and pressure-sensitive skin covering the face



generative nature. The robot uses the internal representation to predict its appearance and the prediction error triggers the movement. However, our model does not yet account for generating the actions directly from prediction errors.

Reaching for the mark is currently engineered. However, we think that this very process of reaching is key to understanding MSR as it is the clearest response and can be analyzed quantitatively. We propose different ways of how

the subjects may reach for the mark (Sect. 3, Fig. 3) and it is our plan to model these in the future. However, more information about how infants and other animals reach for the targets is needed.

Variables that may be instrumental in understanding the mechanisms generating the response are:

- familiarization phase: do the subjects exploit temporal contingency between their movements and those of their reflection in the mirror to test that it is them in the mirror?
- gaze/eye tracking: where do the subjects look: (i) target (mark on the face) in the mirror, (ii) alternate between target and their hand in the mirror, (iii) look at their hand directly
- elements of tactile localization: after an initial inaccurate reaching for the mark, is touch used to bring the hand closer to the target?
- movement duration: from mark placement to touching the mark
- do subjects reach for the correct location but on the other side of their face?
- reaching accuracy
- repeated touches: do the subjects touch the mark repeatedly? Is there any exploration?
- are neck joints/head movements involved? do they assist mark localization or retrieval?
- arm movement kinematics

Additional control experiments could involve: (1) reaching for other targets visible in the mirror or use of distorting mirrors to isolate whether the movements are visually guided on the mirror reflection; (2) reaching to the mark visible in the mirror could be contrasted with reaching for targets that can be perceived by other modalities—in isolation or together with visual perception through the mirror. Chang et al. [15] employed visual-somatosensory stimuli (high-power laser) in macaques; Chinn [16] compared infants' reaching for vibrotactile target on the face away and facing the mirror with the rouge localization task; (3) with infants, one may instruct them also verbally to touch their nose and compare the reaching movements. Finally, one should keep in mind that the mismatch between how one's face normally looks like and the mirror reflection with the mark may not be what is tested. The mark is typically highly salient; additionally, the subject may also interpret the reflection as a conspecific with a mark on the face, which triggers a reaching response to check whether the mark may be also on her own face. One may thus be testing simply reaching for a target on the face with the help of a mirror. Learning the face representation and novelty detection needs also further investigation.

While our own face reflection strongly produces attentional capture, the mark is even more salient. Disambiguation of self-face saliency and pure novelty can be investigated in a MSR setting looking at a non-self face in the mirror and under the face-illusion [45].

7 Conclusion and Future Work

In this work, we provided a mechanistic decomposition, or process model, of what components are required to pass the mirror recognition test. Second, we developed a model on a humanoid robot Nao that passes the test, albeit side-stepping some of the components needed by engineering them. The core of our technical contribution is learning the appearance representation and visual novelty detection by means of learning the generative model of the face with deep auto-encoders and exploiting the prediction error.

The proposed architecture uses a deep neural network to learn the face representation and subsequently deploys this as a novelty detector by exploiting the prediction error. The novelty detection network (autoencoder) is currently based on state of the art in machine learning and computer vision. To what extent this is compatible with the computation in the brain is debatable. Using more biologically realistic neural networks and learning algorithms is a direction of future research. Variance weighted prediction error was relevant to properly detect the mark in the face. Behavioral response—reaching for the mark—was achieved by learning a mapping from the salient region on the face to robot joint configuration required for the reach. The framework was quantitatively tested on synthetic data and in real experiments with different colored marks, sizes and shapes. Furthermore, two robots with completely different visual appearance were used for real-world testing.

Although we were currently investigating how mainly one sensor modality (visual appearance data) produces the behavior, it is important to highlight that self-recognition is a multimodal process [15, 19]. In the future, we hope to acquire additional multimodal data (including movement kinematics and touch) about the details how infants or animals succeed in the mirror mark test and use these as constraints for our computational architecture—adding proprioception and touch. In particular, more information is needed to inform the process of reaching for the mark. One promising computational approach, which directly connects the error prediction scheme of the novelty detection with the action, would be active inference goal driven control [49, 56, 57], where the visual error would produce a reactive reaching behavior toward the mark.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s13218-020-00701-7>) contains supplementary material, which is available to authorized users.

References

- Abrossimoff J, Pitti A, Gaussier P (2018) Visual learning for reaching and body-schema with gain-field networks. In: 2018 Joint IEEE 8th international conference on development and learning and epigenetic robotics (ICDL-EpiRob). IEEE, pp 197–203
- Alzueta E, Melcón M, Jensen O, Capilla A (2019) The ‘Narcissus Effect’: top-down alpha-beta band modulation of face-related brain areas during self-face processing. *NeuroImage* 213:2020
- Alzueta E, Melcón M, Poch C, Capilla A (2019) Is your own face more than a highly familiar face? *Biol Psychol* 142:100–107
- Amsterdam B (1972) Mirror self-image reactions before age two. *Dev Psychobiol* 5(4):297–305
- Anderson JR (1984) The development of self-recognition: a review. *Dev Psychobiol* 17(1):35–49
- Anderson JR, Gallup GG (2015) Mirror self-recognition: a review and critique of attempts to promote and engineer self-recognition in primates. *Primates* 56(4):317–326
- Apps MAJ, Tsakiris M (2014) The free-energy self: a predictive coding account of self-recognition. *Neurosci Biobehav Rev* 41:85–97
- Asada M, Hosoda K, Kuniyoshi Y, Ishiguro H, Inui T, Yoshikawa Y, Ogino M, Yoshida C (2009) Cognitive developmental robotics: a survey. *IEEE Trans Autonon Mental Dev* 1(1):12–34
- Ballard Dana H (1987) Modular learning in neural networks. In *AAAL*, pp 279–284
- Bard KA, Todd BK, Bernier C, Love J, Leavens DA (2006) Self-awareness in human and chimpanzee infants: what is measured and what is meant by the mark and mirror test? *Infancy* 9(2):191–219
- Bigelow AE (1981) The correspondence between self-and image movement as a cue to self-recognition for young children. *J Genet Psychol* 139(1):11–26
- Bringsjord S, Licato J, Govindarajulu NS, Ghosh R, Sen A (2015) Real robots that pass human tests of self-consciousness. In: 2015 24th IEEE international symposium on robot and human interactive communication (RO-MAN). IEEE, pp 498–504
- Bruce V, Young A (1986) Understanding face recognition. *Br J Psychol* 77(3):305–327
- Cangelosi A, Schlesinger M (2015) *Developmental robotics: from babies to robots*. MIT Press, Cambridge
- Chang L, Fang Q, Zhang S, Poo M, Gong N (2015) Mirror-induced self-directed behaviors in rhesus monkeys after visual-somatosensory training. *Curr Biol* 25(2):212–217
- Chinn LK (2019) *Development of Self Knowledge: Tactile Localization to Self-Recognition*. PhD thesis, Tulane University School of Science and Engineering
- Corbetta M, Shulman GL (2002) Control of goal-directed and stimulus-driven attention in the brain. *Nat Rev Neurosci* 3(3):201–215
- Corbetta D, Thurman SL, Wiener RF, Guan Yu, Williams JL (2014) Mapping the feel of the arm with the sight of the object: on the embodied origins of infant reaching. *Front Psychol* 5:576
- de Waal Frans BM (2019) Fish, mirrors, and a gradualist perspective on self-awareness. *PLoS Biol* 17(2)
- De Waal F (2016) *Are we smart enough to know how smart animals are?*. WW Norton & Company, New York
- Diez-Valencia G, Ohashi T, Lanillos P, Cheng G (2019) Sensorimotor learning for artificial body perception. *arXiv preprint arXiv:1901.09792*
- Eimer M (2012) The face-sensitive N170 component of the event-related brain potential. *Oxford Handbook of Face Perception*, pp 329–344
- Fitzpatrick PM, Metta G (2002) Toward manipulation-driven vision. In *Proc. IEEE/RSJ Int. Conf. on intelligent robots and systems*
- Friston K (2010) The free-energy principle: a unified brain theory? *Nat Rev Neurosci* 11(2):127
- Friston KJ, Daunizeau J, Kilner J, Kiebel SJ (2010) Action and behavior: a free-energy formulation. *Biol Cybern* 102(3):227–260
- Fuke S, Ogino M, Asada M (2007) Body image constructed from motor and tactile images with visual information. *Int J Hum Robot* 4:347–364
- Gallagher S (2000) Philosophical conceptions of the self: implications for cognitive science. *Trends Cogn Sci* 4(1):14–21
- Gallup Jr GG (1982) Self-awareness and the emergence of mind in primates. *Am J Primatol* 2(3):237–248
- Gallup GG (1970) Chimpanzees: self-recognition. *Science* 167(3914):86–87
- Gold K, Scassellati B (2009) Using probabilistic reasoning over time to self-recognize. *Robot Autonon Syst* 57(4):384–392
- Guillaume P (1971) Imitation in children. *trans. ep halperin*
- Hafner VV, Loviken P, Villalpando AP, Schillaci G (2020) Prerequisites for an artificial self. *Front Neurobot*:14
- Hart JW (2014) *Robot self-modeling*. Yale University, New Haven
- Heed T, Buchholz VN, Engel AK, Röder B (2015) Tactile remapping: from coordinate transformation to integration in sensorimotor processing. *Trends Cogn Sci* 19(5):251–258
- Hoffmann M, Pfeifer R (2018) Robots as powerful allies for the study of embodied cognition from the bottom up. In: Newen A, de Bruin L, Gallagher S (eds) *The Oxford Handbook 4e Cognition*, chapter 45. Oxford University Press, Oxford, pp 841–862
- Hoffmann M, Marques HG, Arieta AH, Sumioka H, Lungarella M, Pfeifer R (2010) Body schema in robotics: a review. *Autonon Mental Dev IEEE Tran* 2(4):304–324
- Ida Gobbin M, Haxby JV (2007) Neural systems for recognition of familiar faces. *Neuropsychologia* 45(1):32–41
- Kingma DP, Adam JB (2014) A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*
- Kuniyoshi Y (2019) Fusing autonomy and sociality via embodied emergence and development of behaviour and cognition from fetal period. *Philos Trans R Soc B* 374(1771):20180031
- Lafraquière A, Hafner VV (2019) Self-supervised body image acquisition using a deep neural network for sensorimotor prediction. In: 2019 Joint IEEE 9th international conference on development and learning and epigenetic robotics (ICDL-EpiRob). IEEE, pp 117–122
- Lanillos P, Dean-Leon E, Cheng G (2016) Yielding self-perception in robots through sensorimotor contingencies. *IEEE Trans Cogn Dev Syst* 9(2):100–112
- Lanillos P, Dean-Leon E, Cheng G (2017) Enactive self: a study of engineering perspectives to obtain the sensorimotor self through enaction. In: *Joint IEEE Int. Conf. on, in developmental learning and epigenetic robotics*
- Lanillos P, Pages J, Cheng G (2020) Robot self/other distinction: active inference meets neural networks learning in a mirror. In: *European conference on artificial intelligence (ECAI 2020)*
- Latinus M, Taylor MJ (2006) Face processing stages: impact of difficulty and the separation of effects. *Brain Res* 1123(1):179–187
- Ma K, Lippelt DP, Hommel B (2017) Creating virtual-hand and virtual-face illusions to investigate self-representation. *JoVE* 121:e54784
- Mitchell RW (1993) Mental models of mirror-self-recognition: two theories. *New Ideas Psychol* 11(3):295–325

47. Natale L, Orabona F, Metta G, Sandini G (2007) Sensorimotor coordination in a “baby” robot: learning about objects through grasping. *Prog Brain Res* 164:403–424
48. Neisser U (1988) Five kinds of self-knowledge. *Philos Psychol* 1(1):35–59
49. Oliver G, Lanillos P, Cheng G (2021) Active inference body perception and action for humanoid robots. *IEEE Trans Cogn Dev Syst*. <https://doi.org/10.1109/TCDS.2021.3049907>
50. Pablo L, Emmanuel DL, Gordon C (2016) Multisensory object discovery via self-detection and artificial attention. In: *Joint IEEE Int. Conf. on, In developmental learning and epigenetic robotics*
51. Pfeifer R, Bongard JC (2007) *How the body shapes the way we think: a new view of intelligence*. MIT Press, Cambridge
52. Piaget J (1954) *The construction of reality in the child*. Basic Books, New York
53. Pitti A, Mori H, Kouzuma S, Kuniyoshi Y (2009) Contingency perception and agency measure in visuo-motor spiking neural networks. *IEEE Trans Autonom Mental Dev* 1(1):86–97
54. Prescott TJ, Camilleri D (2019) *The synthetic psychology of the self. Cognitive architectures*. Springer, Berlin, pp 85–104
55. Rao RPN, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci* 2(1):79–87
56. Rood T, van Gerven M, Lanillos P (2020) A deep active inference model of the rubber-hand illusion. *International Workshop on Active Inference*. Springer, Cham, pp 84–91
57. Sancaktar C, van Gerven M, Lanillos P (2020) End-to-end pixel-based deep active inference for body perception and action. In: *Joint IEEE 10th international conference on development and learning and epigenetic robotics (ICDL-EpiRob)*
58. Schweinberger SR, Neumann MF (2016) Repetition effects in human ERPs to faces. *Cortex* 80:141–153
59. Steels L, Spranger M (2008) The robot in the mirror. *Connect Sci* 20(4):337–358
60. Sui J, Xiaosi G (2017) Self as object: emerging trends in self research. *Trends Neurosci* 40(11):643–653
61. Tani J (1998) An interpretation of the ‘self’ from the dynamical systems perspective: a constructivist approach. *J Conscious Stud* 5(5–6):516–542
62. Tsakiris M, Longo MR, Haggard P (2010) Having a body versus moving your body: neural signatures of agency and body-ownership. *Neuropsychologia* 48(9):2740–2749
63. Yoshikawa Y, Tsuji Y, Hosoda K, Asada M (2004) Is it my body? body extraction from uninterpreted sensory data based on the invariance of multiple sensory attributes. In: *2004 IEEE/RSJ international conference on intelligent robots and systems (IROS) (IEEE Cat. No. 04CH37566)*, vol 3. IEEE, pp 2325–2330
64. Zaadnoordijk L, Besold TR, Hunnius S (2019) A match does not make a sense: on the sufficiency of the comparator model for explaining the sense of agency. *Neurosci Conscious* 1:niz006
65. Lanillos P, Franklin S, Franklin DW (2020) The predictive brain in action: Involuntary actions reduce body prediction errors. *bioRxiv*. Cold Spring Harbor Laboratory. <https://doi.org/10.1101/2020.07.08.191304>