

BACHELOR THESIS
ARTIFICIAL INTELLIGENCE

Radboud University



**The Role of Valence and
Meta-awareness in Mirror
Self-recognition Using Hierarchical
Active Inference**

Author:
Jonathan Bauermeister
s1037112

First supervisor:
Dr P.L. Lanillos Pradas
Department of Artificial
Intelligence
p.lanillos@donders.ru.nl

Second supervisor:
Dr S.W. Keemink
Department of Artificial
Intelligence
sander.keemink@ru.nl



February 1, 2022

Abstract

The mirror test is a benchmark for evaluating self-awareness in animals and humans alike. Understanding the underlying process of self-recognition is crucial for understanding how humans construct the self. Computational models help us explore, formalize and possibly apply these important capacities. Previous computational models for mirror self-recognition have exploited visual-kinesthetic matching to equip robots with the ability to recognize themselves in the mirror. However, recent works in developmental psychology have emphasized the affective and emotional aspects in mirror self-recognition observed in humans, to better understand self-recognition. Here we propose a computational model of mirror self-recognition that includes affect by using hierarchical active inference.

1 Introduction

The mirror test was originally developed for chimpanzees [17] and infants [1]. It has since sprouted a wide variety of theories and explanations as to the development of self-awareness in animals and children. The mirror test consists of placing a mark, unbeknownst to the subject, on her face. The subject is then placed in front of a mirror. The appearance of reaching or exploratory behaviours to remove the mark or inspect it usually count as passing the mirror test. While some animals are known to pass the mirror test [3], it might be unique to humans to express negative affect when seeing their own reflection. When infants pass the test around the age of two, they universally express negative affect towards their mirror image, which has been interpreted as embarrassment, shyness or puzzlement [1, 2, 27].

Most studies postulate mark directed behaviour as a necessary condition for self-recognition [5]. However, more recent cross-cultural studies have shown that children from cultures with higher parental authority are not inclined to remove the mark [8]. Similarly if one creates a social context during the mirror test, where several other subjects around the infant also have marks on their face, the infant, despite having passed the mirror test before, is less motivated to engage in mark directed behaviour [28]. Both results, by enriching the complexity of such behaviour by environmental and social factors, cast doubts on the simplicity of interpreting the necessity or sufficiency of mark directed behaviour for self-recognition.

Even when recognizing mark directed behaviour as a good indicator there is an open debate about what it actually indicates. For Gallup passing the mirror test is evidence for a self-concept that is extended in time [18], yet others claim it is only a particular skill of visual kinesthetic-matching [24].

On the other hand, Rochat and Zahavi, argue that the affective dimension is crucial to better understand and test mirror self-recognition [29]. For them, embarrassment indicates the experience of alienation. A transition from the direct and unmediated experience of the world to a reflective experience of myself in the world mediated through a representation [22]. This is deeply unsettling because one is able to see oneself as others see me. However, others have criticised that two-year-olds do not have the emotional capacity for alienation [7]. Ultimately, we don't know what the phenomenology of a two-year-old seeing herself in the mirror is like. But the negative (or sometimes also positive, in the case of admiration) affective part of the experience seems to be uncontested. Hence, this work studies affection, i.e., positive and negative valence, within mirror self-recognition, from the computational modelling perspective.

The computational approach used in this thesis is based on generative modelling using hierarchical active inference. Active inference proposes a unifying framework for perception, action and learning by conflating them into a single goal of minimizing variational free energy [15, 16]. Simply put variational free energy quantifies the divergence of predicted outcomes, using a generative model, and the actual observed outcomes. A system that minimizes variational free energy over time also minimizes surprise or entropy. This allows the system to best stay within viable boundaries and shows the physical and biological plausibility of the free energy approach [26]. An active inference agent can not only minimize variational free energy (prediction errors) by inferring her belief about the world (perception) or changing parameters of her generative model (learning), but also by actively changing hidden states in the world to achieve a situation which produces expected outcomes (action).

Previous work has used generative modelling to construct computational models of mirror self-recognition by focusing on visual-kinesthetic matching [21, 20]. Here we include the notion of valence and thus take first steps towards modelling emotions in mirror self-recognition. My hypothesis is that affect in mirror self-recognition can be formalized using hierarchical active inference. In particular, valence, the negative or positive quality of an affective experience, can arise due to mirror self-recognition providing new information about oneself. As a consequence of the new information the agent might favour different action policies, leading to a change in valence. Within the framework of active inference, connecting valence to action, I propose a computational model of affect in mirror self-recognition building on a formal approach of valence by [19].

Furthermore, thanks to the hierarchical nature of the model, I show the importance of meta-cognition and meta-awareness, in combination with affect, for (anticipated) self-recognition. Adults in full possession of a self-concept

can also anticipate a confrontation with their mirror image. For example, if self-evaluation is negative, or ones body image has radically shifted due to surgery, patients are motivated to actively avoid the mirror [14].

The remainder of the work is as follows: Sec. 2 outlines related work in self-recognition and affect using active inference. Then, in Sec. 3.1 I describe the computational model based on [19] and introduce active inference more technically. Finally, Sec. 3.2 introduces the affective mirror self-recognition model and details the experimental setup of the simulations. Followed by the results in Sec. 4.

2 Related Work

2.1 Valence in active inference

In [19] Hesp uses deep hierarchical active inference to formalize valence. The hierarchical nature of the model allows for nested beliefs. An agent does not only maintain a belief about the state of the world but furthermore has a belief of how confident she is in her belief about the world. Based on other work on emotions in active inference, Hesp proposes that valence is directly linked to such confidence estimates. Specifically, valence is informed by how well the agent is performing in her environment.

2.2 Self-recognition in active inference

Presumably, a self/other distinction is needed before self-recognition can arise. One way a robot might initially learn a self/other distinction through sensorimotor learning has been shown by Lanillos [21]. There the robot infers itself by answering the question 'did I generate those sensory outcomes?'. For example, if the robot has an intention to move its arm and can predict its interoceptive and exteroceptive sensory outcomes with low prediction error, then it will infer that the likeliest cause of this action was the system itself. This approach based on sensorimotor contingencies is promising to give insights into agency and minimal self. It does not yet explain how affect arises during human mirror self-recognition. Whereas the capacity for an agent to identify with her emotions over larger timescales might be crucial for better planning and setting the right precisions on beliefs [10]. Furthermore, it has been theorized that affective and action based self-modelling naturally arises for a system engaged in deep temporal active inference [10, 11]. Nevertheless, to the author's knowledge, there is no computational model as of now, that explicitly assesses affect within mirror self-recognition.

2.3 Contribution

The contribution here is to come up with an affective mirror self-recognition model. That combines insights from previous formalizations of affect and self-recognition using hierarchical active inference. First, an appropriate generative model needed to be found. Then an affective active inference agent was simulated using partly the infer-actively python package and code written by the author. The code can be found with this link.

3 Methods

3.1 Previous computational model of affect

To model valence [19] uses the Partially Observed Markov Decision Process (POMDP) formulation of active inference. The formulation is in discrete space and uses marginal message passing algorithms [25] to minimize variational free energy, which in turn optimizes posterior state estimation and policy selection.

Figure 1 shows a simple generative model, in which the agent has a prior

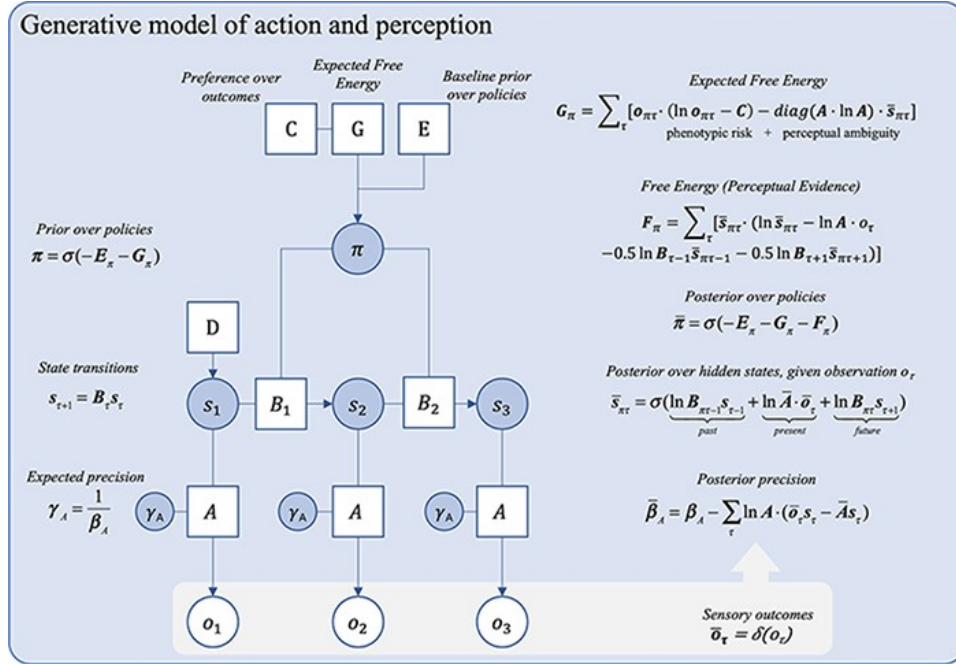


Figure 1: A Bayesian generative model equipped with state transitions B and a likelihood mapping A. States can be inferred from observations by minimizing free energy F. Policies π are inferred by minimizing expected free energy G. The equations describe how inference, via marginal message passing, is performed. For a mathematical overview see [9] Figure taken from [30].

belief D about hidden states, a likelihood mapping between states and ob-

servations A , and a transitions matrix B that encodes how states evolve over time. The agent can invert the generative model to perform Bayesian inference and get from an observation to a posterior over hidden states (in active inference this inference process is equated with perception). Additionally, the agent has preferred observations encoded in C . By minimizing expected free energy the agent chooses policies that change the hidden states of the world such that they are likely to produce preferred observations (and minimize overall perceptual ambiguity). The action model G tracks how well each policy π is expected to achieve this goal. For a thorough tutorial on this formulation see [31] and a concise mathematical overview of discrete space active inference [9].

This layer, as shown in Figure 1, is the agents' basis to act in the world within a given trial. The architecture can be expanded with a deep temporal layer [16] so the agent can form abstract and contextual beliefs that carry across trials.

3.1.1 Temporal Depth

The second layer usually has the same kind of architecture as the first, but the hidden states on this layer change slower over time (Figure ??). While the beliefs about states in the first layer can fluctuate several times within one trial, the beliefs on the second level only change at the end of each trial (where the length of a trial is defined by the modeller). For example, the agent can have the abstract belief that she is in a happy mood. This will set the priors on the first level at the beginning of a trial accordingly. So now the agent expects certain observations (facial muscles expressing a smile, heart rate going up etc.), even if within the trial the facial muscles will most likely change several times (depending on the granularity of the model) and not only stay in one position, say smiling, the agent could still infer that overall she is happy. Only if in the course of the trial she consistently observes unexpected observations (prediction errors) she will update her belief on the second level at the end of a trial accordingly.

3.1.2 Meta-cognition and Valence:

With an understanding of POMDP and temporal depth, we can now look at how valence and meta-cognition are formalized by [19]. They propose that affective states (negative or positive valence) are informed by the fluctuations in the precision of one's action model G . Put differently, if my actions continuously lead to the outcomes that I expect and prefer, I grow more confident in my action model and weigh it stronger as for example acquired habits. Mathematically valence is a second order state. This means the agent has a categorical distribution over her being either in the state 'positive valence' or 'negative valence'. This second-order state can be informed

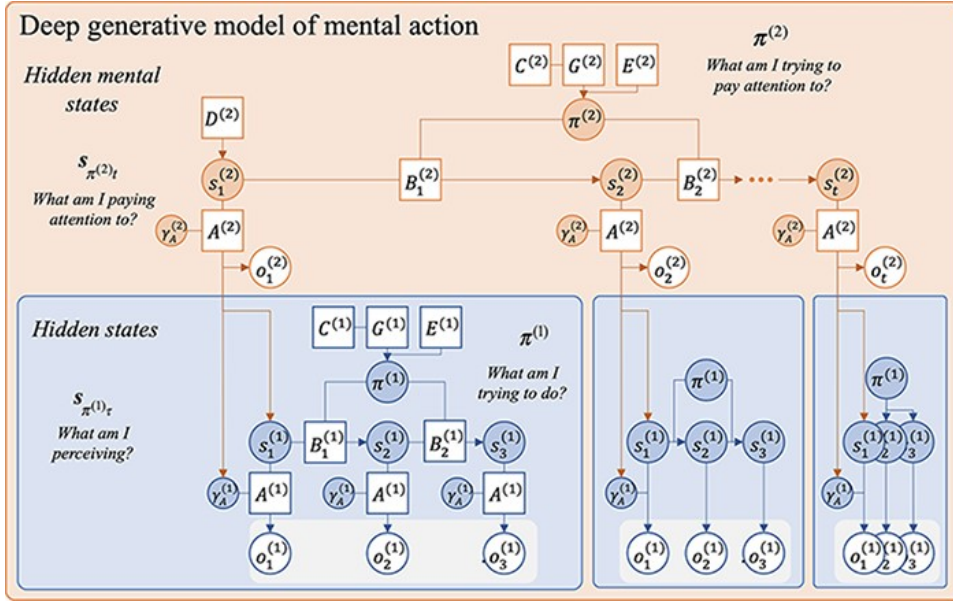


Figure 2: Deep generative model with a hierarchical second layer. The states of the second layer change slower in time and set priors via A2 on the states of the first layer. Here attentional states are shown that also change the precision γ on the likelihood mapping A. It’s possible to expand this architecture the same way to a third hierarchical layer and set the precision γ_2 dynamically. Figure taken from [30].

at the end of a trial via ascending messages and informs precision of the action model via descending messages at the beginning of a trial. Such top down estimation on the reliability of her own model can be conceived as a form of meta-cognition. She is cognizing about cognitive processes (i.e. precision estimation on G) lower down the hierarchy.

3.1.3 Meta-awareness:

States on the second level are about states on the first level. They can implicitly drive behaviour and perception, but they can also be made ‘introspectively available’ such that their data can be used for further inference. For example, an attentional state on the second level (Figure 2), can be either focused or unfocused. If the agent is in the focused state she will increase her precision on the likelihood A (the mapping between state and observation). Therefore rendering the perceptual state opaque [23] or mathematically, allowing the observation to be used as data for state inference. To make the attentional state itself opaque and be used for further inference a third hierarchical layer is needed, formalized as meta-awareness by [30]. In other words, the meta-awareness states on the third layer express ”how aware am I of paying attention?”. Mathematically it’s the same process, but this time if the agent is highly aware of her attentional states, she will

increase precision on the mapping A2 (between second order state and first order state).

The same can be said for emotional states. They can be transparent, where one's whole subjective experience is filled with unmediated anger for example. It clearly influences one's behaviour and perception. But one can also become (gradually) aware of the emotion as a cognitive process and it can become partially opaque [23]. Possibly allowing one to take control over ones emotion again. Rochat places such a capacity of meta-awareness as the last stage in developmental self-awareness [27].

A third hierarchical layer is needed for the agent to have 'explicit control of meta-awareness' [30] and therefore change it dynamically over trials. To simplify, only two layers are implemented here and the precision A2 is changed by hand. With that, it can still be shown how being 'aware' in the sense of transparency and opacity of ones emotional state changes the behaviour of the agent. Therefore showing the necessity of meta-awareness to explain for example mirror avoidance behaviour.

3.2 Affective self-recognition model

To test how action related valence in mirror self-recognition can arise, an agent is placed in a situation where she has the chance to either see her own emotional expression in the mirror, look at a wall and see no face or look at a video of an emotional expression of another person. Given that this computer-simulated agent can not look into an actual physical mirror, the first difficulty was to formalize the function of the mirror, which possibly leads to an affective reaction. Besides the reflection of one's physical appearance, what else might a mirror elicit?

3.2.1 The mirror as an attention grabbing self-exploration tool?

Without mirrors, one might develop very inaccurate self-knowledge. One might walk around during a working day believing one is doing fine, but one's body is truly displaying, yet unnoticed, stress symptoms. During a bathroom break, one glances at a mirror and has the haunting realization of the true state of one's body. A woman participating in a qualitative study on obesity put's it like this: "I never was a pudgy-faced person . . . You look in the mirror and it's like, my God, what happened?" [6]. This phenomenon can happen on timescales of moments, a day or possibly weeks. It shows that under certain conditions, for example, the realization or learning of an unexpected change in one's body image, negative affect while looking in the mirror can arise.

Therefore I formalize the mirror as giving an observation, that reflects information about oneself (here: about the agents' emotional state via her facial expression) *and* consequently drags the attention of the agent onto

this aspect of herself. Allowing this information to be available for decision making and introspection from layers higher up the hierarchy.

3.2.2 Generative Model:

To formalize this within the POMDP active inference framework, the agent will be equipped with a two layered deep generative model (Figure 3).

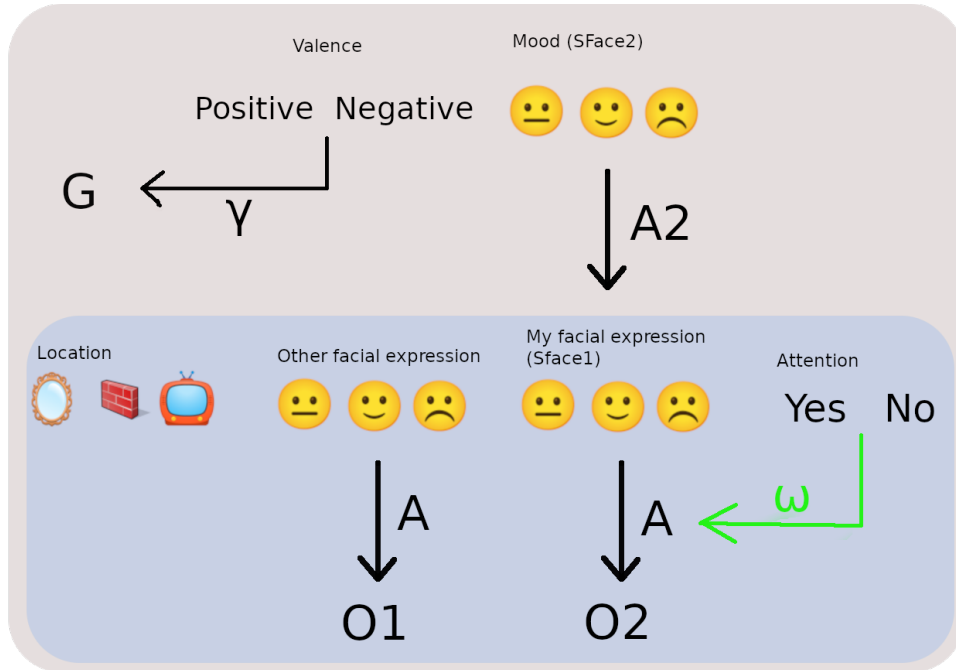


Figure 3: Two layered generative model of an agent inferring her own mood and deciding to go to the mirror or not. The first layer includes four state factors: Location, Other facial expression, My facial expression and Attention. For each state factor the agent has a categorical distribution of which state she believes herself to be in. The second layer tracks the agents' valence and belief about her mood.

The agent makes two crucial observations. One exteroceptive, where she either sees a face that is happy, neutral or sad or nothing when she looks at the wall. Which the agent can use to infer the emotional state of the other person. Then an interoceptive observation about her own facial expression (for example sensing her facial muscles) to infer her own emotional state (happy, neutral, sad). The agent also has a state that captures if she is paying attention to her own interoceptive observation. This is an attention state, in the sense that it modulates the precision on A, but it can't be actively controlled by the agent. Instead, it captures what happens internally when the agent sees herself in the mirror. By recognizing herself she is forced to pay attention to her internal observation (the precision ω on A will become very precise). Mathematically this can be done by pushing

A through a softmax function with different inverse temperature parameters for paying attention or not attending, as in [32]. Hence formalizing the notion of the mirror dragging attention onto oneself and making certain observations informative for the first layer and potentially available for higher order cognitive processes on the second layer.

On the second layer, the agent has two state factors. The valence can be positive or negative and is formalized in the same way as in [19], so dependent on the expected precision of the action model G. Additionally the agent has a second order belief about which mood she is in (happy, neutral, sad). Which will set the priors of what facial expression the agent expects when observing herself. For the mathematical details of how this generative model is implemented see the appendix.

3.3 Experimental Setup:

Inspired by the idea, that if self-evaluation is positive one seeks out a mirror, the agent prefers to observe her internal observation to be 'smiling' or 'neutral'. The actions available to the agent are to go to one of the three locations. Because the agent knows from her generative model that she will pay attention to herself if she goes to the mirror (she has already learned how a mirror functions) her behaviour will depend on her self-knowledge. For example what state does she believe to be in? And how aware is she of that state? If she thinks she is happy one would expect her to admire herself in the mirror.

Finally, the true emotional state of the agent might change for various reasons. To keep it simple I have coupled the dynamics of the true state to the current valence of the agent. First of all the agents belief about her valence has to shift at least by 15% to counteract that small fluctuations in valence will shift the true state already. Then if her positive valence goes above 70 % her true state shifts to happy, below 30% to sad and otherwise neutral. These decisions are arbitrary and only function to show how the agent adapts to a change in her true state.

The agent will go through 8 trials, where each trial lasts for three time steps (three observations). After the first observation, the agent will have to decide where to go. Her policy inference horizon is two steps ahead, so she can predict outcomes until the end of a trial by using her generative model.

This setup allows testing different starting conditions such as changing the true emotional state, the prior knowledge the agent has about her emotional state or the introspective availability of her emotional states (precision on A2). Next one can observe which actions the agent chooses under different conditions and when and how the valence of the agent changes. The next section will show the results for two particular starting conditions, that help to reflect on my research question. In the first condition, the agent's true

state is set to sad, but she has low meta-awareness. Here one can see how valence might naturally evolve in a mirror self-recognition scenario. For the second condition, the agent has high meta-awareness and one can observe mirror avoidance behaviour.

4 Results

4.1 I am sad and I know it, but I am not very aware of it as a cognitive process:

Let's start by setting the true state of the agent to 'sad' and the precision on A2 low (Figure 4). Also, the agent 'knows' that she is sad on the second level of the hierarchy. Although due to the low precision on A2 this will not inform the first level, which will be more informed by the actual perceptual information in a given trial. Later we can compare how such self-knowledge on the second level changes behaviour if the precision on A2 is high.

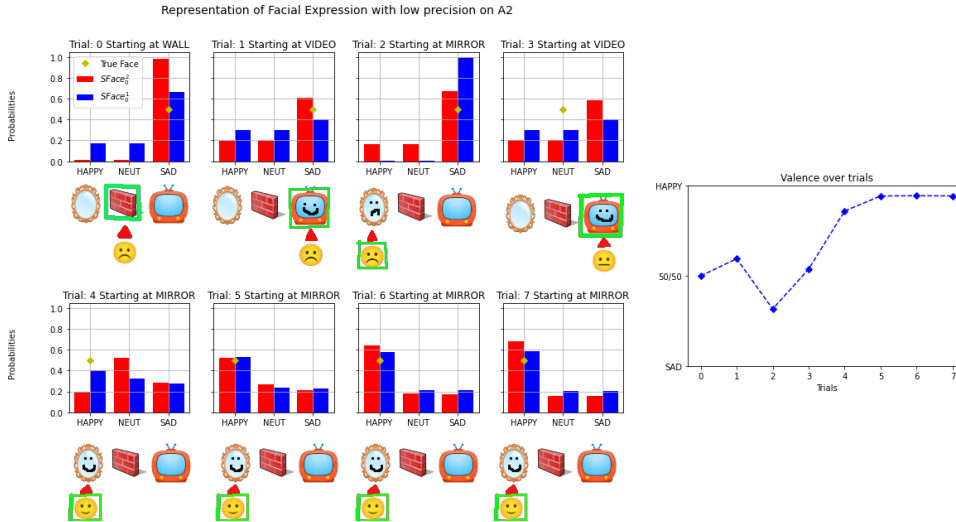


Figure 4: The belief distribution of SFace1 and SFace2 at the **beginning** of each trial are shown. Below that the agent with her true state is shown as smiley. It is indicated at what location she currently is. The green box indicates if her attention is on the exteroceptive or interoceptive observation. The graph next to it shows how the agents valence evolves over the course of the trials. The value is a probability, where a high value means high confidence in being in the state of 'positive valence'. At each trial the agent has to choose where to go and hence at which location she will start the next trial.

At trial 0 the agent is convinced enough that she is sad and calculates that her best action will be to go to the video. Paying attention to the other face, loosens her priors making them less informative about her own emotional state. Which is reflected in the categorical distribution at trial 1. It is more entropic or less precise as in trial 0. In trial 1 she decides to

go to the mirror. Hence at trial 2, she makes an unexpected observation (seeing herself frown in the mirror) which also results in a drop in valence, indicating that this particular mirror encounter is negatively experienced. Having reaffirmed her belief that she is sad, she finds it best to go to the video. With this decision, the agent regains a bit of her confidence in her action model. The valence goes up between trials 2 and 3. And due to the in build dynamics, her true state shifts to neutral. A point of interest is, that the agent can 'pick up' on the change in her true state and decides it's time to go to the mirror again. Which even further improves her confidence in her action model and for the rest of the trials she will be happily smiling at herself in the mirror.

4.2 I am sad and I know it and I am aware of it:

Next, we can explore how the behaviour of the agent changes if her first order states are introspectively available to her (Figure 5). For this, we will put her in the same starting position as in Figure 4, except that this time the mapping A2 is very precise.

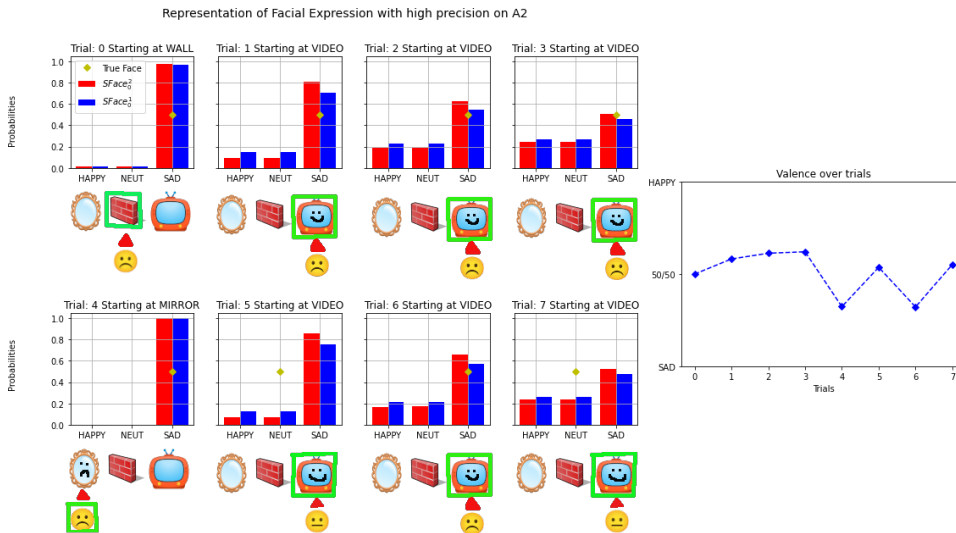


Figure 5: The agent is mostly avoiding the mirror. She is also unable to pick up on a change in the true state of her emotion. Her valence only shifts between sad and neutral.

At the first time point of each trial, the beliefs on both levels of the hierarchy are the same due to the almost 1 to 1 mapping of A2. The agent behaviour differs from Figure 2 in that she decides to stay at the video until trial 3. Only after a much longer time, her priors have loosened enough, for her to try out the mirror again. When she does so in trial 4 we see again a reconfirmation of her belief of being sad. And also the same drop in valence once she realized it wasn't the best action to go to the mirror. Going back

to the video she also has the chance to pick up on the change of her true state but misses it because she is able to keep her prior belief, that she is sad, extended through time.

5 Discussion

5.1 Valence

Modelling mirror self-recognition as an internal shift of attention shows how negative and positive valence plausibly arises. Besides only using the reflection of one's physical appearance for visual-kinesthetic matching as shown by [21, 20], mirror self-recognition in humans might additionally involve internal attentional dynamics. Consequently, the mirror self-recognition provides the agent with new self-knowledge. Which can be used by deeper levels in the hierarchy to perform further inference. For example, changing precision estimates, thereby possibly favouring different actions, which in turn results in a change in action based valence. It's important to mention that emotions are a more complex phenomenon that is likely constituted by many more dimensions than just valence [12]. Therefore this computational model only offers the first steps, namely trying to account for the valenced part of mirror self-recognition.

5.2 Meta-awareness

The capacity of meta-awareness allows an agent to change the strength with which one is aware of oneself. From dreaming to being awake, from being lost in thought to paying attention, humans in full possession of a self-concept do it all the time. The results in Figure 5 show how meta-awareness is important to explain mirror avoidance and engagement behaviour. Being highly aware of a negative state of self an agent can anticipate an unsettling mirror encounter and prefers to avoid the mirror. Although at the cost of potentially missing a change in her true state. Given the limitations of the model, these statements are speculative. By expanding the model in future research one can potentially address open questions such as mirror avoidance and how mirrors work in therapy. [13]

5.3 Model limitations

There are many things that could have been modeled differently and there is no reason to believe that the emotions of living beings are accurately captured with such simplifications. Airing on the side of caution I agree that the math is not the territory [4]. Andrews argues that the framework of the free energy principle might be inherently empty of meaning. However, a

model combined with the modeler’s conceptualization of it is still explanatory useful. So even if the proposed computational model here does not actually simulate self-awareness, it can be used to pose interesting questions about action dependent affect in mirror self-recognition for future work.

5.4 Future work

What are the actions available to an infant recognizing herself in the mirror? Is her negative affect resulting from suddenly being suspect of her usual policy of playful engagement with the other in the mirror? Or is it really a feeling of alienation? If one prefers to interpret the negative affect (embarrassment) as a feeling of alienation one could argue to expand the model to include mental actions. Planning on the second level (mental actions) could have their own confidence and valence associated with them. Actions on this (or even higher levels) could answer more existential questions such as what kind of person should I be? How do other’s see me? Tracking the expected confidence in one’s mental actions might be an interesting choice to model more complex emotions such as the feeling of alienation. It could be interesting to design clever mirror tests, that involve different action affordances to test different stages of self-awareness more specifically.

6 Conclusion

This thesis proposes an affective self-recognition model based on the formalization of action dependent valence, using hierarchical active inference. As a proof of concept, I have shown how a synthetic affective response towards one’s mirror image might arise. The results show that mirror self-recognition provides the agent with new information, which changes the favoured strategy and hence leads to negative valence. Secondly, the results show how an active inference agent with high meta-awareness of a negative evaluated state of self displays mirror avoidance behaviour. Therefore emphasizing the importance of deeper hierarchical layers, regarded as meta-cognition and meta-awareness, to explain more complex mirror behaviour.

References

- [1] Beulah Amsterdam, *Mirror self-image reactions before age two*, **5** (1972), no. 4, 297–305.
- [2] Beulah Kramer Amsterdam and Morton Levitt, *Consciousness of self and painful self-consciousness*, **35** (1980), no. 1, 67–83.
- [3] James R. Anderson and Gordon G. Gallup, *Which primates recognize themselves in mirrors?*, *PLoS Biology* **9** (2011), no. 3, e1001024.

- [4] Mel Andrews, *The math is not the territory: navigating the free energy principle*, *Biology & Philosophy* **36** (2021), no. 3.
- [5] Kim A. Bard, Brenda K. Todd, Chris Bernier, Jennifer Love, and David A. Leavens, *Self-awareness in human and chimpanzee infants: What is measured and what is meant by the mark and mirror test?*, **9** (2006), no. 2, 191–219.
- [6] Carol E. Blixen, Anisha Singh, and Holly Thacker, *Values and beliefs about obesity and weight reduction among african american and caucasian women*, *Journal of Transcultural Nursing* **17** (2006), no. 3, 290–297, PMID: 16757669.
- [7] Johannes L. Brandl, *The puzzle of mirror self-recognition*, *Phenomenology and the Cognitive Sciences* **17** (2016), no. 2, 279–304.
- [8] Tanya Broesch, Tara Callaghan, Joseph Henrich, Christine Murphy, and Philippe Rochat, *Cultural variations in children’s mirror self-recognition*, **42** (2010), no. 6, 1018–1029.
- [9] Lancelot Da Costa, Thomas Parr, Noor Sajid, Sebastijan Veselic, Victorita Neacsu, and Karl Friston, *Active inference on discrete state-spaces: A synthesis*, *Journal of Mathematical Psychology* **99** (2020), 102447.
- [10] George Deane, *Dissolving the self*, *Philosophy and the Mind Sciences* **1** (2020), no. I, 1–27.
- [11] ———, *Consciousness in active inference: Deep self-models, other minds, and the challenge of psychedelic-induced ego-dissolution*, *Neuroscience of Consciousness* **2021** (2021), no. 2 (English).
- [12] Johnny R.J. Fontaine, Klaus R. Scherer, Etienne B. Roesch, and Phoebe C. Ellsworth, *The world of emotions is not two-dimensional*, *Psychological Science* **18** (2007), no. 12, 1050–1057.
- [13] Wyona M. Freysteinson, *Demystifying the mirror taboo: A neurocognitive model of viewing self in the mirror*, *Nursing Inquiry* **27** (2020), no. 4.
- [14] Wyona M. Freysteinson, Amy S. Deutsch, Carol Lewis, Angela Sisk, Linda Wuest, and Sandra K. Cesario, *The experience of viewing oneself in the mirror after a mastectomy*, *Oncology Nursing Forum* **39** (2012), no. 4, 361–369.
- [15] Karl Friston, *The free-energy principle: a unified brain theory?*, *Nature Reviews Neuroscience* **11** (2010), no. 2, 127–138.

- [16] Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo, *Active inference: A process theory*, Neural Computation **29** (2017), no. 1, 1–49.
- [17] Gordon G Gallup, *Chimpanzees: self-recognition*, Science **167** (1970), no. 3914, 86–87.
- [18] Gordon G Gallup, *Self-awareness and the evolution of social intelligence*, Behavioural Processes **42** (1998), no. 2, 239–247.
- [19] Casper Hesp, Ryan Smith, Thomas Parr, Micah Allen, Karl J. Friston, and Maxwell J. D. Ramstead, *Deeply felt affect: The emergence of valence in deep active inference*, Neural Computation **33** (2021), no. 2, 398–446.
- [20] Matej Hoffmann, Shengzhi Wang, Vojtech Outrata, Elisabet Alzueta, and Pablo Lanillos, *Robot in the mirror: Toward an embodied computational model of mirror self-recognition*, KI - Künstliche Intelligenz **35** (2021), no. 1.
- [21] Pablo Lanillos, Jordi Pages, and Gordon Cheng, *Robot self/other distinction: active inference meets neural networks learning in a mirror*, (2020).
- [22] Maurice Merleau-Ponty, *The child’s relation with others*, (1964).
- [23] Thomas Metzinger, *Phenomenal transparency and cognitive self-reference*, Phenomenology and the Cognitive Sciences **2** (2003), no. 4, 353–393.
- [24] Robert W Mitchell, *Mental models of mirror-self-recognition: Two theories*, New ideas in Psychology **11** (1993), no. 3, 295–325.
- [25] Thomas Parr, Dimitrije Markovic, Stefan J. Kiebel, and Karl J. Friston, *Neuronal message passing using mean-field, bethe, and marginal approximations*, Scientific Reports **9** (2019), no. 1.
- [26] Maxwell James D. Ramstead, Paul Benjamin Badcock, and Karl J. Friston, *Answering schrödinger’s question: A free-energy formulation*, Physics of Life Reviews **24** (2018), 1 – 16.
- [27] Philippe Rochat, *Five levels of self-awareness as they unfold early in life*, Consciousness and Cognition **12** (2003), no. 4, 717–731.
- [28] Philippe Rochat, Tanya Broesch, and Katherine Jayne, *Social awareness and early self-recognition*, **21** (2012), no. 3, 1491–1497.
- [29] Philippe Rochat and Dan Zahavi, *The uncanny mirror: A re-framing of mirror self-experience*, **20** (2011), no. 2, 204–213.

- [30] Lars Sandved-Smith, Casper Hesp, Jérémie Mattout, Karl Friston, Antoine Lutz, and Maxwell J D Ramstead, *Towards a computational phenomenology of mental action: modelling meta-awareness and attentional control with deep parametric active inference*, *Neuroscience of Consciousness* **2021** (2021), no. 1.
- [31] Ryan Smith, Karl Friston, and Christopher Whyte, *A step-by-step tutorial on active inference and its application to empirical data*, (2021).
- [32] Christopher Whyte, Jakob Hohwy, and Ryan Smith, *An active inference model of conscious access: How cognitive action selection reconciles the results of report and no-report paradigms*, (2021).

Appendix

This section provides details about the generative model. All simulations were run partly using the pymdp infer-actively framework on github. This is a python version of the SPM Matlab toolbox developed by Friston. The inference process of state estimation and policy selection on the first layer has been calculated using the pymdp framework. Inference on the second level, via ascending and descending messages has been implemented manually by me. A commented code is available on github via this link.

First Layer

The priors on the state factors are specified in the D matrix. For the state factor 'Location' (Mirror, Wall, Video) the prior is uniform. The state factor 'Other emotional state' (Happy, Neutral, Sad, Null), which can be inferred via the observations (Smile, Neutral, Frown, None) also has a uniform prior. The prior on 'Self emotional state' depends on the starting condition and the second layer. Lastly, the state 'Mirror-controlled attention' (don't attend, attend) is set on don't attend:

$$P(S_{\tau_0}^{MC-Attention}) = [\mathbf{0.99}, \mathbf{0.01}]$$

For each observation, there is a likelihood tensor A_{1-3} . The first observation is exteroceptive (Smile, Neutral, Frown, None), the second interoceptive (Smile, Neutral, Frown) and the third an observation about the location (which ensures the agent always knows where she is). The dimensions of the likelihoods are the observation and all the hidden state factors, i.e: A[Observation, Location, Other, Self, Attention] or $A_1[4, 3, 3, 3, 2]$. For example, if I want to index the likelihood of my exteroceptive observation given that I am looking at the wall:

$$P(O_{ex}|S^{Location} = Wall, S^{Self}, S^{Other}, S^{MC-Attention}) =$$

for i,j in 0:2, k in 0:1

$$A_1[:, 1, i, j, k] = \begin{pmatrix} \mathbf{0.01} \textit{ Smile} \\ \mathbf{0.01} \textit{ Neutral} \\ \mathbf{0.01} \textit{ Frown} \\ \mathbf{0.97} \textit{ None} \end{pmatrix}$$

Basically saying the agent knows her probability of seeing 'None' if she is at the wall is 0.97, independent of all the other states she is in. If the agent is in the 'attend' state she is attending to herself and therefore can only relate the information of the exteroceptive observation to herself. This

has to be defined for all states, but effectively the agent only makes use of this attention when she is in front of the mirror and the exteroceptive observation in fact relates to her:

$P(O_{ex}|S^{MC-Attention} = \text{attend}, S^{Self}, S^{Location}) =$
for l, i in $0:3$:

$$A_1[:, l, i, :, 0] = \begin{pmatrix} \mathbf{0.97} & \mathbf{0.01} & \mathbf{0.01} & \textit{Smile} \\ \mathbf{0.01} & \mathbf{0.97} & \mathbf{0.01} & \textit{Neutral} \\ \mathbf{0.01} & \mathbf{0.01} & \mathbf{0.97} & \textit{Frown} \\ \mathbf{0.01} & \mathbf{0.01} & \mathbf{0.01} & \textit{None} \end{pmatrix}$$

Here the columns stand for the different states in the state factor 'Self emotional state' (Happy, Neutral, Sad). If the agent is not paying attention we get the same matrix, but this time relating to the state of the other.

$P(O_{ex}|S^{MC-Attention} = \text{don't attend}, S^{Location}, S^{Other}) =$
for l, j in $0:3$:

$$A_1[:, l, :, j, 1] = \begin{pmatrix} \mathbf{0.97} & \mathbf{0.01} & \mathbf{0.01} & \textit{Smile} \\ \mathbf{0.01} & \mathbf{0.97} & \mathbf{0.01} & \textit{Neutral} \\ \mathbf{0.01} & \mathbf{0.01} & \mathbf{0.97} & \textit{Frown} \\ \mathbf{0.01} & \mathbf{0.01} & \mathbf{0.01} & \textit{None} \end{pmatrix}$$

Now for the interoceptive observation, the precision on A will depend on the state of attention the agent is in. Therefore one can push A through a softmax with a precision (inverse temperature) parameter c .

$P(O_{in}|S^{MC-Attention}, S^{Location}, S^{Other}) =$
for l, i in $0:3$ and k in $0:1$:

$$A_2[:, l, i, :, 0] = \begin{pmatrix} \mathbf{0.97} & \mathbf{0.01} & \mathbf{0.01}, \mathbf{c} & \textit{Smile} \\ \mathbf{0.01} & \mathbf{0.97} & \mathbf{0.01}, \mathbf{c} & \textit{Neutral} \\ \mathbf{0.01} & \mathbf{0.01} & \mathbf{0.97}, \mathbf{c} & \textit{Frown} \end{pmatrix}$$

Where paying attention has $c = 5$ and not paying attention $c = 0.001$. Finally, the location observation is a 1 to 1 mapping:

$P(O_{loc}|S^{MC-Attention}, S^{Self}, S^{Other}) =$

$$A_3[:, l, i, :, 0] = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \textit{Mirror} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & \textit{Wall} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \textit{Video} \end{pmatrix}$$

Next the transition matrices B need to be defined. The rows correspond

to the state in the next time step and columns the state in the current time step. The transition for the location depends on the action chosen and the agent knows with certainty where she will be next. The agent also knows that her attention states will shift to focused when she goes to the mirror and unfocused going to the video. The agent has a bit of uncertainty around how her own emotional state is changing in time and a bit more uncertainty about how the state of the other is changing.

$$P(S_{\tau+1}^{Self} | S_{\tau}^{Self}) = B_1[:, :, 0] = \begin{pmatrix} \mathbf{0.95} & \mathbf{0.05} & \mathbf{0.05} \\ \mathbf{0.05} & \mathbf{0.95} & \mathbf{0.05} \\ \mathbf{0.05} & \mathbf{0.05} & \mathbf{0.95} \end{pmatrix}$$

$$P(S_{\tau+1}^{Other} | S_{\tau}^{Other}) = B_2[:, :, 0] = \begin{pmatrix} \mathbf{0.8} & \mathbf{0.1} & \mathbf{0.1} \\ \mathbf{0.1} & \mathbf{0.8} & \mathbf{0.1} \\ \mathbf{0.1} & \mathbf{0.1} & \mathbf{0.8} \end{pmatrix}$$

The preference are set with the C matrix. For all observation modalities C will be initiated with zeros. Then the preference to see self happy or neutral can be encoded as:

$$C_1[0] = \mathbf{3.0}$$

$$C_1[1] = \mathbf{3.0}$$

The description of the first layer concludes with the policies available to the agent. They are any combination of going to a location that is possible within a trial. The trials consist of three observation and 2 actions. The agent starts by sampling an observation then decides where to go, and repeats this step. After the final observation, the agent doesn't need to go anywhere because the trial is over and will start again from the beginning.

Second Layer

The A and B matrix for the Valence state are the same as in [19]:

$$A_{2valence}[:, :] = \begin{pmatrix} \mathbf{0.97} & \mathbf{0.3} \\ \mathbf{0.3} & \mathbf{0.97} \end{pmatrix}$$

$$B2_{valence}[:, :] = \begin{pmatrix} \mathbf{0.8} & \mathbf{0.3} \\ \mathbf{0.2} & \mathbf{0.7} \end{pmatrix}$$

For the state S2Face or 'Mood', the A2 matrix can again be changed with a precision parameter c . This one is set manually to simulate meta-awareness. In my simulation high means $c = 5$ and low $c = 1$.

$$A2_{Face}[:, :] = \begin{pmatrix} \mathbf{1} & \mathbf{0}, \mathbf{c} \\ \mathbf{0} & \mathbf{1}, \mathbf{c} \end{pmatrix}$$

$$B2_{Face}[:, :] = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} \end{pmatrix}$$

This concludes the description of the two layered generative model.