Towards Conscious RL Agents By Construction

Asen Nachkov
INSAIT, Sofia University
Sofia, Bulgaria
asen.nachkov@insait.ai

Abstract

The nature of consciousness has been a long-debated concept related to human cognition and self-understanding. As AI systems become more capable and autonomous, it is an increasingly pressing matter whether they can be called conscious. In line with narrative-based theories, here we present a simple but concrete computational criterion for consciousness grounded in the querying of a virtual self-representation. We adopt a reinforcement learning (RL) setting and implement these ideas in SubjectZero, a planning-based deep RL agent which has an explicit virtual self-model and whose architecture draws similarities to multiple prominent consciousness theories. Being able to self-localize, simulate the world, and model its own internal state, it can support a primitive virtual narrative, the quality of which depends on the number of abstractions that the underlying generative model sustains. Task performance still ultimately depends on the modeling capabilities of the agent where intelligence, understood simply as the ability to model complicated relationships, is what matters.

Keywords: artificial intelligence, consciousness, reinforcement learning

Introduction

Artificial intelligence and cognitive science loosely form a *primal-dual* problem pair. Both are progressing at great speed, but there still remains a duality gap, at the heart of which stands artificial consciousness and the question of how experiences could arise in a formal computational setting. Despite multiple promising theories (Dennett, 1993; Clark, 1998; Pfeifer & Bongard, 2006; Tononi, 2004; Lamme, 2006; Baars, 1993), capturing the subjective nature of *what it's like* (Nagel, 1980) has proved very difficult so far.

To make progress, we intentionally adopt an approach of *ultra*-simplification. In what follows, we first build a shared understanding of how subjective experiences can be reduced to *variable processing* and how the self can be reduced to a virtual representation of a look-alike agent similar to us. Subsequently, we present an algorithm designed to embody the essential characteristics necessary for having experiences. At all times the guiding principles in our reasoning are *conceptual simplicity* and *concreteness*.

Agency. Agency is the ability to exert purposeful change in the environment. Each one of us can be viewed according to two different perspectives – first, as an individual with a given personality (Allport, 1937) – and second, as a biological agent capable of adapting to an environment and harvesting rewards (Sutton & Barto, 2018). Let us call the first perspective the humanistic one, and the second perspective the reinforcement learning one.

The **RL perspective** belongs to the real physical world, where there are clear-cut limits on what is possible and what is not. It emphasizes an environment in which we are reward-maximizing agents. And the rewards which we maximize are, at least in their very basis, biochemical (Schultz, Dayan, & Montague, 1997). It is only through our biochemistry that we perceive various states of the world as pleasant or unpleasant, desirable or undesirable (Rolls, 2000). Our base reward functions are hard-coded by evolutionary processes and we rarely can change them willingly (Cosmides & Tooby, 1994).

The **humanistic perspective** is based on the view of a person – an individual with inherent qualities, values and worth. This representation does not entail agency and exists only inside the mental model of the world constructed by our brains (Johnson-Laird, 1983). There are no limits to the world states here as we can imagine or believe anything. Our identities, memories, and anticipations exist only in this virtual abstract space, likely with the purpose of giving us a better representation for dealing with long-term tasks requiring extended planning in the present (Oyserman, Elmore, & Smith, 2012).

A generative model. It is *useful* to think of this virtual space as a generative predictive model of the world around us (Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Clark, 2013), *irrespective of whether the brain actually works like that*. We can generate future representations of the world – anticipations – or past representations from sparse signals – memories. Our world models are flexible and efficient: they can condition their representations on physical signals or virtual ones (Friston, 2010). We can associate states of the world to biochemical rewards (Berridge & Robinson, 1998), as well as physical sensory triggers like vision, hearing, and smell to imaginations and memories.

The virtual self. Inside one's world model, there is a representation for a person, who looks like them, talks like them, and behaves like them. This representation is built from multiple signals gathered through time: reflections in the mirror (Rochat, 2003), correlations between personal actions and subsequently obtained biochemical rewards (Schultz et al., 1997), and feedback from other agents. Natural language is a communication interface to these representations (Chomsky, 2002; Pinker, 1994). During our early development, each one of us labels these features "I", "me", or "self", at which point

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

one can refer to a person who looks like them and behaves like them. We call this an identity in this paper – *an explicit virtual representation of a physical agent who is identical in appearance, preferences, and behaviour to us*.

It is our belief that the self-awareness in each and every one of us is as simple as described. One needs to recognize themself first in the third person perspective, as their own doppelgänger, and only then can those learned features be renamed. Just like the word "chair" refers to the mental representation of a chair, so does the word "I" refer to the mental representation of a particular human exactly like us.

Consciousness. Using this relation between the physical agent and their virtual representation, we can provide a *computational* definition for consciousness. *Consciousness is the process of continuously associating the sensory information processed by an agent in the physical world to their virtual "self" representation*. The association can be understood as the querying, or conditioning on, or projecting sensory features onto the virtual representation. Self-awareness, understood as only having access to an explicit virtual "self" model, is a necessary but insufficient condition for consciousness.

Minimalist nature. This approach to handling consciousness is minimalist, but concrete enough to be implementable in practice. It abstracts away any relation to the sensor modalities, autonomy, task selection, and even intelligence. The latter of these is understood simply as the ability to model complicated relationships. Thus, intelligence is only crucial to develop *accurate and realistic* self-representations.

Contributions. We highlight that the reasoning above represents a pragmatic, implementable, and *simplified* computational model for consciousness. Driven by the desire to assess its utility, our main contribution is the implementation and evaluation of a digital agent which puts this model to practice. We do not claim novelty for the ideas established in this paper, though we have taken great care in presenting them in a way that emphasizes the physical-virtual nature, which we believe is key to understanding consciousness.

Related Work

From the cognitive science side, prominent theories of consciousness like **recurrent processing theory** (Lamme, 2006, 2010, 2020), **global workspace theory** (Baars, 1993, 1997), and **higher-order theories** (Lau & Rosenthal, 2011; Lycan, 2001; Brown, Lau, & LeDoux, 2019) focus on the neuroscientific and computational processes underpinning consciousness – whether it is by recurrent processing, by a global finite information store, or by representations about representations. They do not address how these computational aspects relate to the feeling of what it's like. The **attention schema theory** (Graziano, 2017) claims that the brain builds a map of its own attention just like it builds a map of the body. While useful in explaining self-attention, this still does not explain the content of our perceptions. To some extent this is addressed by the **sensorimotor theory** (O'regan & Noë, 2001), which claims that our experiences are formed by attending to

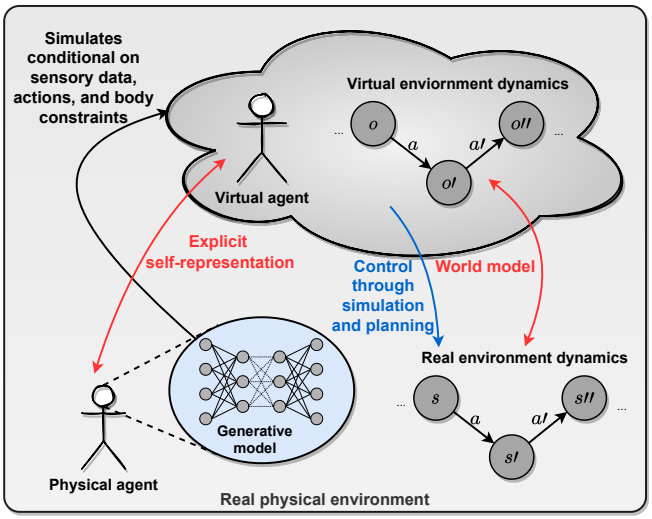

Figure 1: **A schematic of our approach.** A physical agent simulates a virtual model of the environment, projects the sensory information to the virtual agent and plans its actions.

specific contingencies relating the sensory input and our action outputs. Further, **embodied cognition theories** (Varela, Thompson, & Rosch, 2017), assert that the experiences of an agent are dependent on and reflect also the biological constraints of the body and the environment.

The above theories provide a fairly good understanding of the real information in an experience, and what computational mechanisms are needed to effect it. But our experiences are continuous, temporally-correlated, and thus merge into "narratives". According to Dennett (1993), there are **multiple narrative drafts** consistent with one's sensory experience and they constantly compete with each other, prompting some of them to fade out or dominate our consciousness. The **cortical conductor theory** (Bach, 2019) provides similar reasoning. Overall, this view of consciousness as a narrative fits well with the more general **predictive processing** belief (Friston, 2010) and resonates deeply with the generative modeling and the humanistic view presented above. Our work is grounded in that context. We cannot comment on other theories, such as IIT (Tononi, 2004, 2012), which do not easily offer similar generative explanations. We also do not make any claims as to which anatomical regions are responsible for producing experiences (Merker, 2007; Crick & Koch, 1990).

On the machine learning side, RL methods like Muzero (Schrittwieser et al., 2020), Perceiver (Jaegle, Gimeno, et al., 2021; Jaegle, Borgeaud, et al., 2021), Dreamer (Hafner, Lillicrap, Ba, & Norouzi, 2019), and AdA (Team et al., 2023) all focus on important aspects – planning, attention, world models, adaptability to new tasks – but do not have an *explicit* virtual self-model, synchronized to the acting agent. Schmidhuber (1991) claims that due to the feedback present in the world model, self-introspective capabilities can be observed. We find this only natural, albeit still reactive to the agent's own "reflection". This aspect is present also in LLMs where even though they can plan their own verbal outputs

(Wei et al., 2022; Yao et al., 2023), any self-modeling arising from the autoregressive token generation is still implicit. To address this issue, we believe the agent needs to have explicit semantic information of themself and the capability to self-identify – something we implement in our solution.

It is not uncommon to explicitly engineer different priors within the agent's behaviour. Curiosity and intrinsic motivation have been explored in great depth as ways to deal with sparse rewards and to learn meaningful skills even in the presence of no task-specific supervision (Barto, 2013; Eysenbach, Gupta, Ibarz, & Levine, 2018; Pathak, Agrawal, Efros, & Darrell, 2017; Aubret, Matignon, & Hassas, 2019; Burda, Edwards, Pathak, et al., 2018; Burda, Edwards, Storkey, & Klimov, 2018). Similarly, risk-sensitive agents (Dabney, Ostrovski, Silver, & Munos, 2018) can prioritize actions favourable to their risk-seeking profile. Nonetheless, while having inductive biases resembling curiosity and risk aversion, these agents do not have the semantic understanding to recognize or interpret them as such.

Overall, we believe that cognitive theories *need to be formalized* in a computational setting to make them more concrete, while in the machine learning contexts one *needs to engineer explicit self-modeling*. The resulting model will speak the language of both fields, bridging the gap between them. This is what we attempt in this work.

Environment, Body, and Experiences

The virtual agent is embodied in a physical agent and thus has to account also for the system complexity of the physical body. This requires access to sensors and actuators.

Sensors. Humans have many sensors for processing information external to the body, including photo-, mechano-, and thermoreceptors. They convert physical quantities like light photons, pressure, and temperature into electric impulses. But we also have baro- and chemoreceptors for interoception. Thus, similar to how we recognize objects external to the body, we can recognize states internal to it (Craig, 2002).

Actuators. They convert electric impulses into muscle contractions for locomotion in the external environment and chemical secretions for regulation in the internal one. We do not have awareness over them as such low-level process attention may be superfluous for high-level decision making.

Neurons. The action potentials of neurons can be interpreted as both data features to be processed and instructions for other cells. Depending on the wiring, neural networks can act as generators, selectors, modulators, or feature extractors. Likewise, a semantic classifier alone does not distinguish whether the input neural spikes represent real sensory data, e.g. an object we are looking at, or have been generated by another neural module, such as when we are imagining the object. This aspect, combined with recurrence, allows for a digital agent to simulate different futures or pasts.

Subjective experiences in humans serve an information processing role. It is useful to distinguish hot from cold, or hunger from thirst because these experiences are crucial for survival. And the underlying information processing is more *variable*, owing to the unique connections in each individual's cortex, rather than subjective. The subjective aspect likely results from the brain constructing an interpretation of the temporally correlated sensory features which constitutes the story in which the virtual agent exists (Seth, 2013).

We consider a simple example. Suppose a bear is charging at you. The sudden novel data from the external environment enter the brain and activate the autonomic nervous system, triggering a fight-or-flight response and raising one's pulse and breathing rate. In turn, interoceptors detect the increased pressure, providing information about the body's reaction to the outside event. Thus, *real* information both external and internal to the body is available for processing by the cortex.

It is by processing this real data that "subjective" experiences are formed. By predicting and generating features corresponding to colour, spatial positions, tactile sensations, auditory frequency, semantics, all conditioned on the raw sensory data, the brain effectively reconstructs the agent's environment, along with itself in it (Friston, 2010). Since the processed features are saved into memory, they can be recalled in the future. We highlight that the agent's subjective experiences are part of its world model and, therefore, must be themselves generated. Thus, the agent *believes* that it feels, where beliefs can be understood simply as state features with high probability under one's own generative world model. Recalling these features from memory simply reconstructs the story in which the agent "feels". As a consequence of these mechanics, the resulting world interpretations can be considered intrinsic, private, and ineffable. Their existence can be verified only if the agent reports so.

Implementation

Here we build a planning-based RL agent which has an explicit learned self-representation on which its behaviour is conditioned. We consider the Atari Arcade Learning Environment (Bellemare, Naddaf, Veness, & Bowling, 2013) – a staple benchmark for classic RL agents like DQN (Mnih et al., 2015), C51 (Bellemare, Dabney, & Munos, 2017), and MuZero (Schrittwieser et al., 2020). We showcase our prototype approach on Pong. The other environments are left for the future when we scale up the agent. We call our agent SubjectZero, because it is based on MuZero (Schrittwieser et al., 2020) while having explicit self-representation.

We consider the visual observations from the environment to represent a flat world in which the agent lives. They also contain sufficient information for the agent to observe itself – a necessary condition in order to learn a self-representation. The reward function, according to which rewards are generated, is fixed and exogenous to the virtual self. Having the agent come up with tasks is possible by sampling different reward functions but is not crucial for our setup so we do not discuss it in details. Since the self-representation abstracts the data modality, we believe the Atari setting is a good candidate for a simple environment on which to experiment.

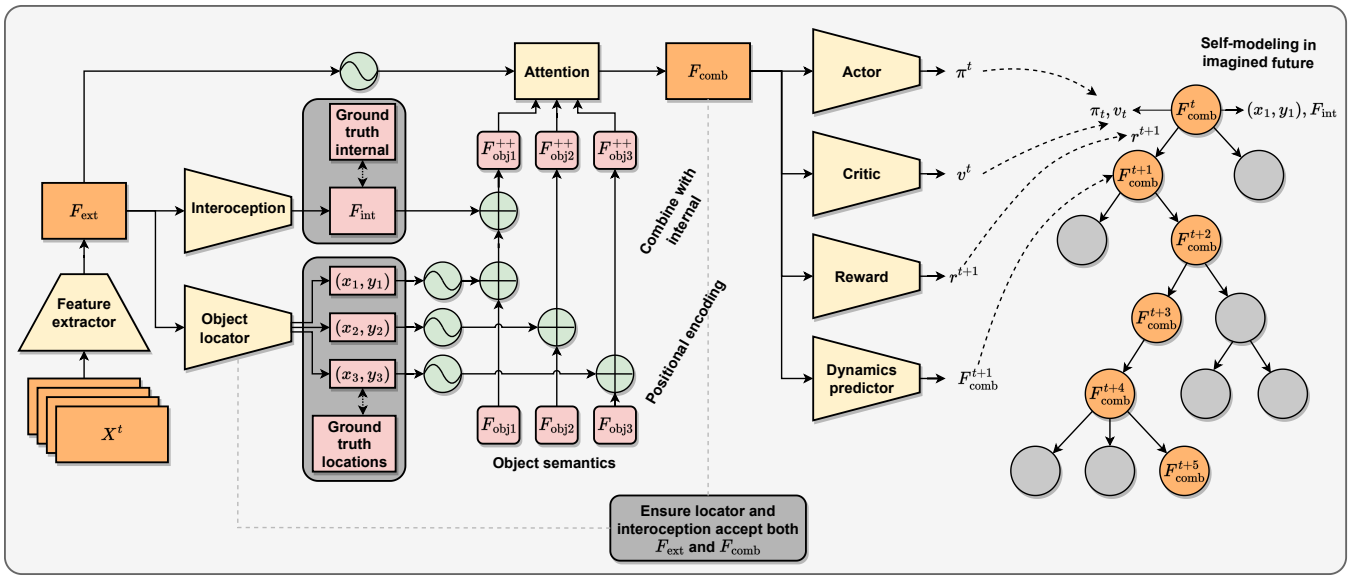

Figure 2: **Architecture of SubjectZero**. A convolutional feature extractor obtains high level features F_{ext} from the visual sensory inputs. An object locator predicts object coordinates from those features. They are trained in a supervised manner with the ground truth locations. The predicted object coordinates are encoded and added to a number of learned object representations F_{obj}. An interoceptive component produces features F_{int} which are supervised by the reward function, or a similar signal modeling the agent's body. They are added to the object corresponding to the agent itself. Subsequently, an attention block aggregates all available features, producing F_{comb}. From it, an actor, critic, reward, and dynamics network branch out. Their outputs are used by the Monte Carlo Tree Search (MCTS), similar to MuZero. Yellow components are learnable networks. Red components are explicit and interpretable semantic objects meaningful in the world model. Orange objects are latent signals which are not interpretable but are useful for processing. Most time indices for the current step are omitted for clarity.

Architecture. Figure 2 shows the proposed architecture. The agent receives temporally stacked grayscaled visual frames as input. A **convolutional feature extractor**, roughly corresponding to brain areas V_1, V_2, and V_5 (for motion), extracts a low-dimensional latent representation F_{ext} from the sensory observations. These features correspond to the real sensory information external to the physical agent.

We model the agent's body by introducing a custom environment-specific function $\tilde{r}$ that returns a scalar value *similar* to the reward. This function is fixed by the algorithm designer and represents all the immutable feedback loops and self-regulating mechanisms that the body possesses. Its output is a numeric label, representing the body's reaction to outside events, and for Pong it is based on whether the ball is beyond the position of any of the players:

$$\tilde{r} = \begin{cases} -1, & \text{if } x_{\text{ball}} > x_{\text{player}} \\ 1, & \text{if } x_{\text{ball}} < x_{\text{opponent}} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Crucially, by predicting this numeric label using a neural network, the agent obtains latent features F_{int}, corresponding to the interoceptive assessment of its own body's state. We purposefully choose $\tilde{r}$ to be very similar to the actual reward function r to allow the agent to condition any downstream processing on this early "estimated reward". This is uncommon in model-based RL, where the estimated reward is treated as an output, only ever used for the final action-selection, not for conditioning intermediate representations.

The latent state F_{ext} is passed to an **object locator**, a small fully-connected network which outputs (x, y) coordinates for each of a number of learnable objects, as well as confidences α for whether they are present in the current state. The coordinates represent the pixel locations in the coordinate system of the visual input image. While a human can effortlessly model the surrounding real 3D geometry, our Atari agent lives in these flat images and for it the only geometry that matters is that of the pixels. The object locator corresponds to the dorsal visual stream in the brain.

We endow the agent with the capacity to represent a limited number of **objects or concepts explicitly**. The semantics of these objects are represented by vectors F_{obj}, learned using gradient descent. Since they are not in the network weights, they are explicit representations. Such vectors are similar to token embeddings in NLP, learnable camera poses in SLAM (Sucar, Liu, Ortiz, & Davison, 2021; Z. Zhu et al., 2022), and object queries in DETR-like models (Carion et al., 2020; X. Zhu et al., 2020). In the case of Pong, we use just 3 representations corresponding to the two paddles and the ball. Once trained, these representations remain fixed at test time. They are roughly equivalent to explicit semantic memory in humans. Explicit episodic memory is given by the replay buffer and implicit procedural memory is provided by the weights of the policy network.

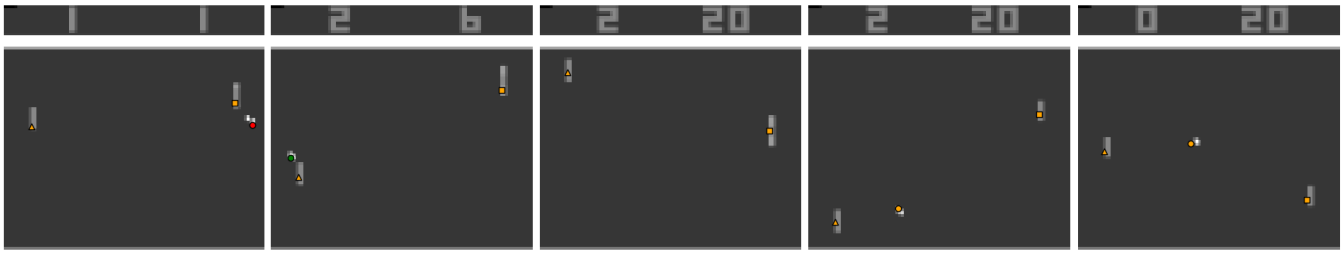

Figure 3: **Predictions**. We showcase samples from different test runs. The player is the right paddle. Scores are shown at the top. In all cases the agent accurately predicts the object visibilities and locations. The ball is colored green, red, or orange based on the agent's own prediction of whether the current observed state is beneficial, harmful, or neutral, as given by the interoception module. Overall, the agent learns to play Pong effortlessly while at the same time being able to self-localize.

Subsequently, the agent uses **learnable positional encodings** to embed the predicted locations to the object semantics, according to the estimated visibility, obtaining features F^{+}_{obj}. This operation fuses positional and semantic information, in effect representing the object at that particular location as a single vector. The environment features F_{ext} are encoded similarly to yield F^{+}_{ext}. We add the interoceptive features F_{int} to one of the vectors in F^{+}_{obj} – that which represents the agent itself, obtaining F^{++}_{obj}. The resulting vectors can be thought of as **objects in context**. At this point, the features F_{ext} represent the current state, as informed by the agent's sensors, while F^{++}_{obj} - the objects along with their locations and semantics.

We combine them by using a simplified cross-attention block (Vaswani et al., 2017) where the queries are the object vectors and the lookup keys are the state features F^{+}_{ext}. We compute the attention coefficients without further aggregation, and use them to modulate the addition of F^{+}_{ext} and F^{++}_{obj}:

$$F_{\text{comb}} = F^{+}_{\text{ext}} + \sigma(F^{+}_{\text{obj}}(F^{+}_{\text{ext}})^{\mathsf{T}})F^{++}_{\text{obj}}. \tag{2}$$

Here, $\sigma(\cdot)$ is a softmax activation. In essence, this operation adds the semantic information for the objects into only those spatial positions from the state feature map which are relevant for that object. This produces enhanced features F_{comb}, containing the semantic object information at the right spatial locations. All computations here preserve spatial relationships, similarly to the retinotopic map in V_1 and, overall, this part corresponds to the posterior parietal cortex, which summarizes the spatial object relations in the scene.

From the object-focused latent state F_{comb} we have standard branching **policy**, **value**, **reward**, and **state dynamics** networks. These components allow the agent to simulate future state dynamics and plan its actions accordingly. We follow the discrete planning approach of MuZero (Schrittwieser et al., 2020) and utilize Monte Carlo tree search (MCTS) for the action-selection. The planning and policy modules corresponds to the prefrontal cortex. The reward and value predictors are related to the ventral striatum in the human brain.

We also utilize an additional **self-supervision** loss, as in Ye, Liu, Kurutach, Abbeel, and Gao (2021), which forces states close to each other to also have similarly close representations. This component is used only for practical reasons to increase the sample efficiency of our method.

Training. Our implementation is based on the LightZero library (Niu et al., 2023). We train the localization and interoceptive networks using supervised learning. The ground truth labels for the object locations are obtained from the game's underlying RAM state. The mapping from raw bytes to pixel locations was estimated manually by us and is environment-dependent. For a more general treatment of unsupervised object discovery, where it is difficult to obtain ground truth annotations, we recognize that more sophisticated methods are needed (Locatello et al., 2020).

Future simulation. Our agent can accurately localize the objects, as shown in Figure 3. But we can also compose the locator with the dynamics predictor, to localize objects in imagined future trajectories. To that end, since the dynamics predictor returns the next F_{comb}, we train the locator and the interoception network to produce accurate coordinates from both F_{ext} and F_{comb}. This allows the agent to produce meaningful object tracklets from within its own imaginations, as shown in Figure 4. MuZero can also construct virtual trajectories, but they are represented as dense latent features. Compared to them, SubjectZero can apply in the simulated future any network that assigns meaning to the hidden vectors.

Properties

As a consequence of its architecture and design, the agent proposed above has the following properties:

1. It can self-localize. By training it to predict the locations of objects, one of which is itself, it learns to predict its own location from its own visual appearance.
2. It can track objects across sequences of frames. This results from the technical fact that when training, predictions are matched to ground truth annotations always by index, as opposed to e.g., by total distance, as in Carion et al. (2020).
3. It can represent its own body's state. By predicting and conditioning downstream calculations on features constructed from the body itself, it learns how the collected rewards and subsequent actions relate to its interoceptive and proprioceptive data. This is similar to the sensorimotor theory (O'regan & Noë, 2001).
4. The representation of its virtual-self is explicit and is stored exactly in one of the learned vectors F_{obj}. By predicting

real pixel coordinates this representation is also *grounded* to the real physical environment, not the one in its dynamics predictor. Since the self-representation is used to compile information from its surroundings, the agent essentially conducts a query on its virtual self.

5. Related to the recurrent processing theory (Lamme, 2006), the dynamics predictor is recurrent and forms the backbone for the simulated rollout needed for action selection. The overall process is *control through simulation*. To take an action, the agent self-localizes, combines static knowledge with the dynamic surroundings, simulates possible futures and selects the action yielding the most beneficial future.
6. In the attention block, the agent's attention is spread out over the features. But if we use multiple attention heads, this would create different *narratives* for what the agent attends over. If they are then merged, the resulting bottleneck will force the different narratives to compete, which is similar to the multiple drafts theory (Dennett, 1993).
7. By disabling the attention to some of the features, different altered states of consciousness can be modeled. Removing the explicit object localization will make the agent consider the effects of other objects only implicitly, as if only by reaction and instinct. Removing the interoceptive features will prevent any aspects of the body to ever be considered.

Access consciousness. In principle, the attention weights used for aggregating information can themselves become the ground-truth targets of another network that will predict them from sensory data. This will allow the agent to predict its own attention, similar to the attention schema theory (Graziano, 2017). Furthermore, using the predicted attention features as additional inputs to the dynamics predictor, policy, or any other component can be interpreted as access consciousness (Block, 1995), i.e., making the agent's own attention accessible to the generative story. We do not implement it here because the predicted attention map does not contain information related to the task itself and is unneeded. Perhaps in a multi-task setting, where the only information that transfers between different tasks is that related to the body, explicit modeling of the agent's own attention will be invaluable.

Phenomenal consciousness. We highlight that the presented agent naturally does not have humanlike experiences. However, by accurately modeling its own body's state, and by being able to self-localize, it has enough functionality to construct a primitive virtual narrative in which it exists and feels. In fact, in this narrative existence is the *default*, because of how the architecture has been constructed. Thus, the presented setup is enough to represent a virtual self F_{obj1}, located at (x, y) within the scene F_{ext}, with preferences given by the reward function, and current experiences F_{int}. With more such abstractions, the narrative will only become richer and more convincing. Verbal explanations can then be built on top of these features to report on them and communicate. The resulting agent will be constrained by its generative world model to believe that it exists, without understanding why and how.

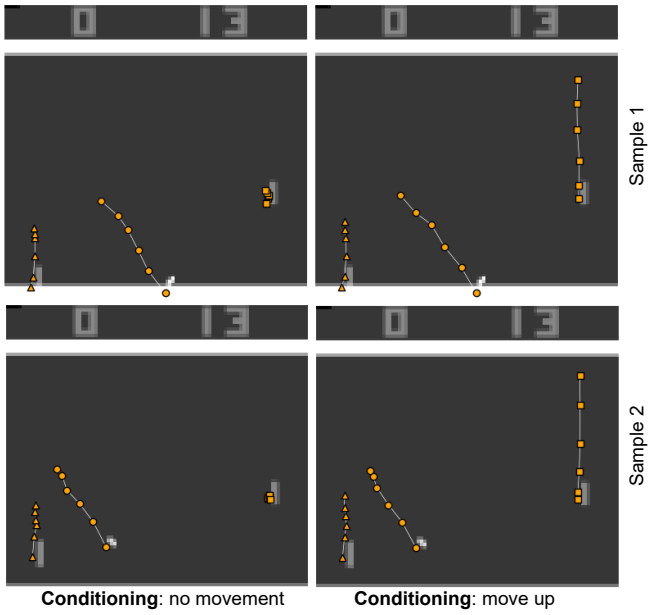

Figure 4: **Imagined trajectories**. Our agent can track objects in the imagined future, conditional on its own actions. The rows show two different samples at test time. The columns show two different action sequences - one in which the agent does not move, and one in which it moves up. The imagined trajectories are consistent with these actions.

Conclusion

In this work we have projected the problem of consciousness from the abstract opaque philosophical setting into a concrete and verifiable reinforcement learning one. We presented a simple high-level functional definition of consciousness based on the querying of a virtual self-representation and have described SubjectZero, a planning RL agent, which incorporates explicit self-modeling by construction and has many desirable properties and characteristics, reminiscent of those believed to govern consciousness in humans.

Compared to methods specialized in reasoning about object permanence (Traub et al., 2022), event segmentation (Gumbsch, Adam, Elsner, Martius, & Butz, 2022), or planning using the learned model (Hafner, Lillicrap, Norouzi, & Ba, 2020), our approach focuses on establishing the minimal elements needed to build a virtual story in which the agent lives. This generative story allows us to talk about subjective experiences in a formal computational setting, thereby providing an improved understanding of artificial cognition. In general, we anticipate that combining deep learning with cognitive architectures will be a fruitful and *meaningful* endeavour and will open up the doors to many full stack cognitive engineering solutions in the future.

The practical benefits of such a line of work will be improved interaction with AI systems, more interpretable and grounded decision-making, better alignment with goals, tasks, or desires. On the conceptual side, we hope this approach leads to further insights in these fields.

References

Allport, G. W. (1937). Personality: A psychological interpretation.

Aubret, A., Matignon, L., & Hassas, S. (2019). A survey on intrinsic motivation in reinforcement learning. *arXiv preprint arXiv:1908.06976*.

Baars, B. J. (1993). *A cognitive theory of consciousness*. Cambridge University Press.

Baars, B. J. (1997). In the theatre of consciousness. global workspace theory, a rigorous scientific theory of consciousness. *Journal of consciousness Studies*, *4*(4), 292–309.

Bach, J. (2019). The cortical conductor theory: Towards addressing consciousness in ai models. In *Biologically inspired cognitive architectures 2018: Proceedings of the ninth annual meeting of the bica society* (pp. 16–26).

Barto, A. G. (2013). Intrinsic motivation and reinforcement learning. *Intrinsically motivated learning in natural and artificial systems*, 17–47.

Bellemare, M. G., Dabney, W., & Munos, R. (2017). A distributional perspective on reinforcement learning. In *International conference on machine learning* (pp. 449–458).

Bellemare, M. G., Naddaf, Y., Veness, J., & Bowling, M. (2013). The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, *47*, 253–279.

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain research reviews*, *28*(3), 309–369.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and brain sciences*, *18*(2), 227–247.

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the higher-order approach to consciousness. *Trends in cognitive sciences*, *23*(9), 754–768.

Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., & Efros, A. A. (2018). Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*.

Burda, Y., Edwards, H., Storkey, A., & Klimov, O. (2018). Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213–229).

Chomsky, N. (2002). *Syntactic structures*. Mouton de Gruyter.

Clark, A. (1998). *Being there: Putting brain, body, and world together again*. MIT press.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, *36*(3), 181–204.

Cosmides, L., & Tooby, J. (1994). *Origins of domain specificity: The evolution of functional organization*. na.

Craig, A. D. (2002). How do you feel? interoception: the sense of the physiological condition of the body. *Nature reviews neuroscience*, *3*(8), 655–666.

Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. In *Seminars in the neurosciences* (Vol. 2, pp. 263–275).

Dabney, W., Ostrovski, G., Silver, D., & Munos, R. (2018). Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning* (pp. 1096–1105).

Dennett, D. C. (1993). *Consciousness explained*. Penguin uk.

Eysenbach, B., Gupta, A., Ibarz, J., & Levine, S. (2018). Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, *11*(2), 127–138.

Graziano, M. S. (2017). The attention schema theory: A foundation for engineering artificial consciousness. *Frontiers in Robotics and AI*, *4*, 60.

Gumbsch, C., Adam, M., Elsner, B., Martius, G., & Butz, M. V. (2022). Developing hierarchical anticipations via neural network-based event segmentation. In *2022 ieee international conference on development and learning (icdl)* (pp. 1–8).

Hafner, D., Lillicrap, T., Ba, J., & Norouzi, M. (2019). Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*.

Hafner, D., Lillicrap, T., Norouzi, M., & Ba, J. (2020). Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*.

Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., ... others (2021). Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*.

Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., & Carreira, J. (2021). Perceiver: General perception with iterative attention. In *International conference on machine learning* (pp. 4651–4664).

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness* (No. 6). Harvard University Press.

Lamme, V. A. (2006). Towards a true neural stance on consciousness. *Trends in cognitive sciences*, *10*(11), 494–501.

Lamme, V. A. (2010). How neuroscience will change our view on consciousness. *Cognitive neuroscience*, *1*(3), 204–220.

Lamme, V. A. (2020). Visual functions generating conscious seeing. *Frontiers in Psychology*, *11*, 83.

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in cognitive sciences*, *15*(8), 365–373.

Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., ... Kipf, T. (2020). Object-centric learning with slot attention. *Advances in Neural*

Information Processing Systems, *33*, 11525–11538.

Lycan, W. G. (2001). A simple argument for a higher-order representation theory of consciousness. *Analysis*, *61*(1), 3–4.

Merker, B. (2007). Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and brain sciences*, *30*(1), 63–81.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... others (2015). Human-level control through deep reinforcement learning. *nature*, *518*(7540), 529–533.

Nagel, T. (1980). What is it like to be a bat? In *The language and thought series* (pp. 159–168). Harvard University Press.

Niu, Y., Pu, Y., Yang, Z., Li, X., Zhou, T., Ren, J., ... Liu, Y. (2023). *Lightzero: A unified benchmark for monte carlo tree search in general sequential decision scenarios.*

O'regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, *24*(5), 939–973.

Oyserman, D., Elmore, K., & Smith, G. (2012). Self, self-concept, and identity.

Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning* (pp. 2778–2787).

Pfeifer, R., & Bongard, J. (2006). *How the body shapes the way we think: a new view of intelligence*. MIT press.

Pinker, S. (1994). The language instinct: How the mind creates. *Language. New York: Harper Collins*.

Rochat, P. (2003). Five levels of self-awareness as they unfold early in life. *Consciousness and cognition*, *12*(4), 717–731.

Rolls, E. T. (2000). On the brain and emotion. *Behavioral and brain sciences*, *23*(2), 219–228.

Schmidhuber, J. (1991). A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats* (pp. 222–227).

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., ... others (2020). Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, *588*(7839), 604–609.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in cognitive sciences*, *17*(11), 565–573.

Sucar, E., Liu, S., Ortiz, J., & Davison, A. J. (2021). imap: Implicit mapping and positioning in real-time. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 6229–6238).

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Team, A. A., Bauer, J., Baumli, K., Baveja, S., Behbahani, F., Bhoopchand, A., ... others (2023). Human-timescale adaptation in an open-ended task space. *arXiv preprint arXiv:2301.07608*.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, *331*(6022), 1279–1285.

Tononi, G. (2004). An information integration theory of consciousness. *BMC neuroscience*, *5*, 1–22.

Tononi, G. (2012). The integrated information theory of consciousness: an updated account. *Archives italiennes de biologie*, *150*(2/3), 56–90.

Traub, M., Otte, S., Menge, T., Karlbauer, M., Thuemmel, J., & Butz, M. V. (2022). Learning what and where: Disentangling location and identity tracking without supervision. *arXiv preprint arXiv:2205.13349*.

Varela, F. J., Thompson, E., & Rosch, E. (2017). *The embodied mind, revised edition: Cognitive science and human experience*. MIT press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... others (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Ye, W., Liu, S., Kurutach, T., Abbeel, P., & Gao, Y. (2021). Mastering atari games with limited data. *Advances in Neural Information Processing Systems*, *34*, 25476–25488.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2020). Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.

Zhu, Z., Peng, S., Larsson, V., Xu, W., Bao, H., Cui, Z., ... Pollefeys, M. (2022). Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 12786–12796).