Implementing Self Models Through Joint-Embedding Predictive Architecture

Yuyue Jiang (yuyuejiang@ucsb.edu)
University of California, Santa Barbara, CA 93106 USA

Dezhi Luo (ihzedoul@umich.edu)
University of Michigan, Ann Arbor, MI 48109 USA

Abstract

Self models contribute to key functional domains of human intelligence that are not yet presented in today's artificial intelligence. One important aspect of human problem-solving involves the use of conceptual self-knowledge to detect self-relevant information presented in the environment, which guides the subsequent retrieval of autobiographical memories that are relevant to the task at hand. This process enables each human to behave self-consistently in our own way across complex situations, manifested as self-interest and trait-like characteristics. In this paper, we outline a computational framework that implements the conceptual aspect of human self models through a modified version of the joint-embedding predictive architecture. We propose that through the incorporation of human-like autobiographical memory retrieval and self-importance evaluation, the modified architecture could support machine agents with significantly enhanced self-consistency, which could be applied to deliver more believable simulations of human behaviors.

Keywords: self models, JEPA, autobiographical memory, conceptual self, machine self consistency, cognitive AI

Introduction

Humans are able to learn vast amounts of knowledge about the world in relatively small exposure and yet know how to apply them to reason and act in novel situations. This general intelligence (Hassabis et al., 2017; Summerfield 2022) has been attributed to our ability to develop and use abstract mental models of the world, or world models (Ha & Schmidhuber, 2018; Friston et al., 2021), to form predictions of future world states and plan for actions accordingly. Recently there are attempts to implement world models into machine agents as a pathway to achieve human-like general intelligence (Eslami et al., 2018; Ha & Schmidhuber, 2018; Schrittwieser et al., 2020; Assran et al., 2023; Rao, Gklezakos, & Sathish, 2023), which have succeeded in out-performing corresponding specialized programs for a wide range of tasks in several domains.

On the other hand, there are important aspects of human intelligence that cannot be captured by world models alone. One of such aspects that has been largely overlooked in artificial intelligence research is the self. While the term's use in cognitive science is wide-ranging, in this current paper we specifically focus on the self as the collection of perceptual and conceptual information that records one's individual experiences, thoughts, and actions (Kihlstrom et al., 1988). The presence of these mental models of self, or self models (Vogeley et al., 1999; Northoff, 2013), enables humans to reason and act with impressive self-consistency. On one hand, humans maintain and pursue complex, long-term goals that are contingent upon our personal beliefs and values, which is often described in terms of the possession of self-interest (Moore & Loewenstein, 2004). On the other hand, during these goal-oriented behaviors, humans exhibit characteristics that vary significantly across individuals but consistent among oneselves, often referred to as traits (Matthew, Deary, & Whiteman, 2003). Self-interest and trait-like characteristics are two functionally significant features of human intelligence that are not presented in today's artificial intelligence but are vital to its functional improvement and public acceptance (Pelau, Dabija, & Ene, 2021).

We propose the computational simulation of human self models through the extension of the joint-embedding predictive architecture (JEPA; as in LeCun, 2022) as an approach to enable human-like self-consistency in machine performance. Designed to support intelligent agents capable of solving domain-general tasks, JEPA centers around a predictive world model module that simulates the general pattern of higher cognition in humans without considering individual differences shaped by experiences. Under the rationale of cognition-inspired artificial intelligence (Cassenti, Veksler, & Ritter, 2022), we suggest that with modifications of its key components, JEPA could be extended to encompass the information-processing framework of self models to personalize its performance. Designs of these modifications are formulated by considering the mapping between relevant modules in the JEPA framework and the neurocognitive substrate of self models in humans, which we discuss below.

Human Self Models

Overview

Human self models can be said to involve two key components: (1) autobiographical memory and (2) conceptual self-knowledge. Autobiographical memory is further differentiated into perceptual and conceptual kinds. The retrieval of conceptual autobiographical memories serves an important role in humans' self-consistent decision-making across complex situations and is guided by the processing of conceptual self-knowledge. We therefore suggest that formalizing these conceptual self models is significant to our present goal.

Autobiographical Memory

A particular life event that happened to a human agent can be remembered both perceptually and conceptually, which are processed through distinctive neural networks (Tulving,

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.*

1984; Brown et al., 2018). These two facets of autobiographical memory form the basis of the autobiographical memory system (Neisser, 1986; Conway & Pleydell-Pearce, 2000; Conway, 2005), which plays an important role in one's judgment and decision-making in both close and open-ended tasks (Simon et al., 1987; Sheldon, Fenerci, & Gurguryan, 2019).

Perceptual Autobiographical Memory Memories of perceptual details during moment-specific events constitute the perceptual aspect of the autobiographical memory system. These perceptual autobiographical memories are critical references for humans in solving close-ended tasks, given that their solutions are contingent upon the context of the problems. Detection of perceptual cues from the surroundings activates the posterior hippocampus to retrieve perceptual memories of past events that are relevant to the present tasks, which is then served as a case-by-case template for decision-making (Sheldon, Fenerci, & Gurguryan, 2019). While perceptual autobiographical memory is an important constituent of the human self models and is reliable in solving close-ended tasks, we do not mean to explore its computational adaptation in our model due to both functional and feasibility reasons. To begin with, perceptual memory is less relevant in solving open-ended tasks, given that in these cases perceptual details do not contain cues for accessing task-relevant memory segments, which is the central concern of this project. Furthermore, reinstating and applying perceptual memory requires conscious access, which machine adaptation faces significant engineering challenges and ethical concerns (Krauss & Maier, 2020).

Conceptual Autobiographical Memory In contrast to perceptual autobiographical memory, the conceptual aspect of the autobiographical memory system consists of both episodic memories and semantic ones. Conceptual episodic memories (i.e. conceptual details of discrete, moment-specific events) decay rapidly after formation (Talamini & Gorree; 2012). However, elements of each discrete memory during their consolidation process contributes to the thematic, knowledge-like memories of one's life over a longer time scale (e.g. life stories and general events), often referred to as autobiographical semantic memories or autobiographical knowledge (Conway & Pleydell-Pearce, 2000; Conway, 2005), which could be retained for extended periods of time, even after corresponding discrete memories have been forgotten. Autobiographical knowledge is suggested to be hierarchically organized in terms of abstraction level and stored across the brain (Conway, Singer, & Tagini, 2004; Prebble, Addis, & Tippett, 2012), constituting the memory system known as the autobiographical knowledge base (Conway & Pleydell-Pearce, 2000). Given that solutions to open-ended problems are not contingent to contexts, retrieval of autobiographical knowledge with task-relevant conceptual information is critical for one's decision-making in complex social situations (Conway, Singer, & Tagini, 2004; Conway, 2005; Sheldon, McAndrews, & Moscovitch, 2011). We therefore suggest that the implementation of conceptual autobiographical memory could be a significant addition to cognitive-inspired artificial intelligent systems for achieving human-like performance in open-ended tasks. In particular, such performances are marked by a significant degree of self-consistency, which is enabled by conceptual self-knowledge in its role among the retrieval of autobiographical knowledge, which we discuss below.

Conceptual Self-knowledge

The active employment of contents from the autobiographical knowledge base (Conway & Pleydell-Pearce, 2000) during goal-oriented processing contributes to the construction of a separate system of self-relevant information often referred to as the conceptual self (Neisser, 1988; Conway, Singer, & Tagini, 2004; Demiray & Bluck; 2011). This separate system contains a rich collection of conceptual self-knowledge not limited to relational self-schema, personal beliefs and values, and long-term goals (Kihlstrom & Cantor, 1984; Klein & Loftus, 1993; Conway, Singer, & Tagini, 2004). Sustained by neural networks separate from that of autobiographical knowledge base (Grilli & Verfaellie; 2015), the conceptual self is instrumental in the evaluation and retrieval of task-relevant autobiographical knowledge for decision-making, and is especially responsible for self-consistency across performances.

Specifically, the cue-detection process that guides the retrieval of task-relevant memory segments from the autobiographical knowledge base is enabled by the schematic evaluative processes mediated by the medial prefrontal cortex (mPFC), which indexes conceptual self-knowledge to assess the self-importance of specific conceptual information presented in the task environment and activates autobiographical knowledge containing corresponding cues (Hampton, Bossaerts, & O'Doherty, 2006; D'Argembeau, 2013; Vaidya & Badre, 2020; Levorsen et al., 2023). Specific segments of autobiographical knowledge are retrieved to inform performance not only based on its relevance to the tasks, but also whether it has high self-importance according to one's conceptual self-knowledge, such as relational self-schema, personal beliefs and values, and long-term goals (D'Argembeau, 2013). The conceptual self therefore has been referred to as the underlying representation of self-interest and trait-like characteristics (Kihlstrom & Cantor, 1984; Klein & Loftus, 1993), which essentially enable one's performance in complex, novel situations to be consistently aligned with past experiences. In that respect, we propose that the computational implementation of the conceptual self along with its interactive mechanisms with the autobiographical knowledge base is critical for machine agents to achieve human-level self-consistency. Notably, since conceptual self-knowledge guides autobiographical knowledge retrieval by serving as the schematic input of the evaluation network centered in the mPFC, we suggest that the implementation of the conceptual self for decision-making could be reduced to a simulation of the self-importance evaluation mechanism, without any construction of individual conceptual self-knowledge.

The Original JEPA

Overview

Joint-Embedding Predictive Architecture (JEPA) is a cognitive architecture trained with self-supervised learning with the implementation of world models processing that closely resembles human cognition (LeCun, 2022). The perception module represents the current state of the world, taking in aspects of reality and sending them to the world model module, which predicts potential future world states based on imagined action sequences proposed by the actor. The cost module, comprising the intrinsic cost and the trainable critic, computes a single scalar output referred to as "energy," measuring the agent's discomfort level. The actor module computes a chain of actions that may optimally respond to the world. The operation of JEPA begins with the perception system generating a representation of the current external world state *sx*(0) = P(*x*), with the cost module simultaneously computing the immediate cost associated with the state. Following, the actor proposes an initial sequence of actions (*a*(0), . . . , *a*(*t*),…, *a*(*T*)). Given the proposed action sequence, the world model then predicts likely world state representations which is again fed in the cost module for estimation of the total cost, represented as F(*x*) = C(*sx*(*t*)). The actor then proposes a new action sequence with a lower cost within several iterations. Once a low-cost action sequence is converged upon, the actor sends the actions in series to the configurators to implement. The entire cycle repeats for the next perception-action episode.

Latent Variable Z & Energy-based Model Training

The distinction between JEPA and most predictive models lies in their approach of evaluating the prediction as compared to the outcome. Instead of minimizing the divergence between predictions calculated from inputs *sx* and actual outcomes *y*, JEPA compares the prediction based on inputs *sx* with the prediction of outcomes based on the actual outcome (a perception of *y*, hence *sy*). This selective processing process aligns closely to human cognition, in the sense that we do not need to process every element of the external world to make choices but instead relies on a perception of the world that filters out non-salient details. However, this auto-determined selective processing of all information dimensions may lead to the collapse of the system due to uncontrolled minimization of perceptual dimensions to maximize efficiency. To prevent this, JEPA introduces a latent variable Z to the predictor in addition to *sx*. The content of Z at a specific timeframe may come from the training dataset or from aspects not included by the perception module in x itself, thus also providing additional dimensions alongside the perception *sx* to aid prediction.

Due to JEPA's structure being incompatible with the probabilistic modeling training used in traditional machine learning, LeCun (2022) employed an approach based on an implicit energy function F during JEPA training known as the energy-based model. F represents the compatibility between *x* and *y*. Specifically, when a pair (*xi*, *yi*) exhibits high compatibility, the energy function takes a low value, and vice versa. F can therefore capture the dependence between *x* and *y*. During training, the main optimization goal is formalized as follow:

$$\check{z} = \underset{z\in\mathcal{Z}}{\operatorname{argmin}} E_w(x,y,z) = \underset{z\in\mathcal{Z}}{\operatorname{argmin}} D\big(s_y, \operatorname{Pred}(s_x,z)\big)$$
$$F_w(x,y) = \min_{z\in\mathcal{Z}} E_w(x,y,z) = D\big(s_y, \operatorname{Pred}(s_x,\check{z})\big)$$

Namely, we want the pair (*x*, *y*) to have minimal energy, which indicates higher compatibility and more accurate prediction. Given this motivation, optimizations are specified in order to maximize the information content of sx from x and sy from y, facilitating prediction of sy from sx and also minimize the information content of the latent variable z used in the prediction.

Cost Module

The cost module assesses the discomfort of the agents and represents it using it as an internal energy state. A lower energy level is associated with less discomfort and vice versa. The Cost module comprises two parts: (1) the intrinsic cost module, corresponding to human's innately positive or negative values attributed to basic biological states such as pleasure and pain; (2) the critic, which is a trainable and optimizing module that takes current intrinsic cost as an input to predict potential values of intrinsic cost in the future. The prediction of the critic is trained through access to short-term associative memory. It is noteworthy that the design and training of the cost module is a completely data-driven process through a single projection from the short-term associative memory with Markovian property without any externally registered rules, therefore minimizing bias from human assumptions.

Implementing Self Models Into JEPA

Memory Module

In the original JEPA, z represents a reservoir of information derived from both the current stimuli that isn't included in perception (i.e. the unattended aspects of stimuli) and a broader pool of information not present in the original world (e.g. the entire training set). This module is necessary for preventing system collapse while improving predictions of the world state by serving alongside the information in the perceptual module. However, since it is based entirely on undesignated inputs, the specific content of Z in a given timeframe is completely unpredictable. Therefore, while the inclusion of Z improves predictions for each individual task, it also creates additional inconsistency for the agent's behaviors across situations. On the contrary, as illustrated in previous sections, humans apply relevant autobiographical knowledge through self-importance evaluation as an aid to perception to improve predictions. The way conceptual components of self models are applied in this process resembles that of z in the original JEPA, but unlike the latter it drastically improves the consistency of one's action across

timeframe by providing self-generated input based on one's past experiences.

We therefore propose an alternative to the original latent variable Z by introducing a memory module simulating the processing of conceptual self models in humans. The memory module consists of a Long-Term Memory (LTM) submodule including autobiographical knowledge base and discrete task memories, and a Conceptual Self submodule which functions as a classifier that models after the role of the mPFC in self-importance evaluation. This memory module takes the perception of current situation (*sx*) or prediction (pred(*sx*)) to be processed under the classifier, which compares them with a set of labels that were each automatically created based on their respective category of discrete conceptual memories. The design of the memory module is described below in detail.

Label Generation The LTM system maintains a collection of discrete task memories, which each structured as:

$$M\big(s(x)_i, \text{pred}\,\big(s(x)_i, a(s(x)_i)\big), a(s(x)_i), \text{cost}\,\big)$$

A label of a category is determined as the average of all memory content it includes, which encompass all dimensions of its 4 components: *sx*, Pred, *a*, and Cost. *sx*'s dimensionality is determined by the perception module， pred's dimensions is based on the perception and the optimal actions series, a is the list of actions based on the predictions made by the world model, while the total cost is represented as a singular numerical value, with detailed computation elucidated in the following section. Notably, all components stored here are in their final state that is outside the iteration of the modules and thus encompassing the final predictions, the optimal actions series and the final cost. Within the LTM submodule, a neural network categories all existing discrete memories based on the similarity of the perception *sx*, resulting in an indeterminate number of categories through unsupervised learning. This categorization prioritizes maximizing homogeneity within categories and heterogeneity between them, without relying on predesignated rules. An essential difference between the simulation and human cognition is that here we refer to discrete memories as lower-level autobiographical knowledge as opposed to episodic conceptual memories, as for machine decision-making there is no need to represent events in its non-abstracted, declarative form due to the absence of awareness. Additionally, autobiographical knowledge base in the LTM submodule refers to only the set of the highest-level autobiographical knowledge marked as labels as opposed to the entire hierarchy of autobiographical knowledge.

Classification Rather than exhaustively running through all discrete memories for ones that matches with the current perception, the Conceptual Self submodule scans the perception for matching elements with the set of labels and from which activates corresponding categories in the autobiographical knowledge base, a process simulating the mPFC self-

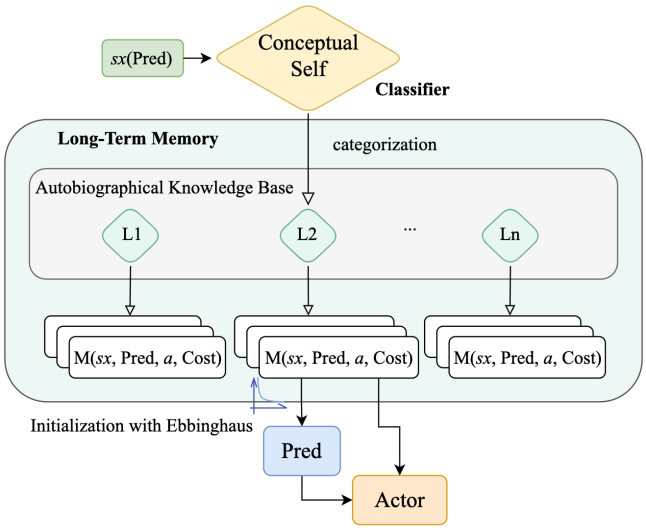

Figure 1: The memory module As a novel perception (*sx*) or prediction (pred(*sx*)) emerges, it is incorporated into the system and compared against the labels of all categories (L1, L2, …, Ln), leading to its classification into a specific category (L2 in the graph). Subsequently, these labels act as cues for retrieving pertinent memories. The system then returns the original memories within the designated category, along with their respective prediction and corresponding weight values (as indicated by the Ebbinghaus curve on the retrieval arrow), to the world model module (pred) and actor module respectively, enhancing their predictive capabilities.

importance evaluation process. This activation allows the memory contents in the LTM within the paired category to be accessed and retrieved, thus enabling the involvement of more intricate and specific underlying memories in the ongoing process for predictions in world model module and actor module, contributing to improved predictions. Following the prediction phase, the newly acquired memories are stored in LTM to be available for future reference in subsequent events.

Memory Initialization Despite being abstracted into a time-invariant knowledge, lower-level conceptual memories aren't constantly accessible due to the process of forgetting, as in human cognition. Obtaining a label doesn't assure the successful retrieval of all related memories since the activation of lower-level conceptual memories is significantly time-dependent (Davis & Zhong, 2018). We thus specifically consider using the Ebbinghaus forgetting curve to initialize weight assignments (Ebbinghaus, 1964). The most recent memory is assigned a weight of 100%, while other memories have weights that progressively decrease based on their entry time. We employ a time stamp framework that shifts at the entering of new events. This approach also prevents the algorithm of categorization to fall into local optima, ensuring the identification of a global optimum based on memory availability. The function of activation of the neural network is as following:

$$f(\sum(w * \bar{x}) + b)$$

where the synaptic weights are represented as:

$$w = \frac{100k}{\log(t)^c + K}$$

Recollection and Reparameterization This process enables the assimilation and adjustment of categories (labels) with the addition of each new discrete task memory. When a new perception recalls memories linked to a specific category, these memories are consolidated as the result of reactivation (Schiller & Phelps, 2011). This reactivation, achieved by incorporating them into recent memories, refreshes the accessibility of the memories. In the model, this indicates that the memory enters a new forgetting curve, resulting in an augmentation of its weight value. The mechanism for this augmentation involves setting the weight of the most recent memory (initially with the highest weight) to 100%. For the remaining memories, their weights are adapted based on the highest memory curve. For example, if the highest memory in the initially extracted category was at 80% and increases to 100%, this 125% increment is applied to all memories. Their weight values are multiplied by 125%, demonstrating the adjustment across all retrieved memories.

Dimensionality Reduction Replacing latent variables with autobiographical memory in the JEPA framework necessitates additional processes for dimensionality reduction. Since reducing dimensionality before storing memory might result in a significant loss of information, we suggest that it could instead be done during memory retrieval. Specifically, we propose that autoencoder is a suitable technique for dimensionality reductions in the given framework, provided that it can efficiently handle complex, high-dimensional data and maintain the information richness of temporality and activation frequency by adjusting hidden layers and network structures.

Personalized Cost Module

Cost in the original JEPA model is based entirely on intrinsic cost, which only reflect hardwired values like the inherently negative valence of hunger and pain. This approach thus provides a universal metric for all value judgment and is suitable for supporting agents having identical values across all domains which are determined by a set of given fixed rules. Human value judgment, however, is characterized by remarkable individual differences. This is because it is supported by a hierarchical and multifaceted cost-value system, in which intrinsic values like basic biological drives are only one among several crucial factors (Maslow, 1954; Kenrick et al., 2010). The rest, ranging from socially constituted needs to self-actualizations, are deeply grounded in one's past experiences and are represented in the Conceptual Self. In the likely cases in which different levels of needs clash with each other, higher-order cognitive functions make use of the Conceptual Self to make a personalized value judgment, a feature not evident in JEPA due to the sole reliance on intrinsic cost.

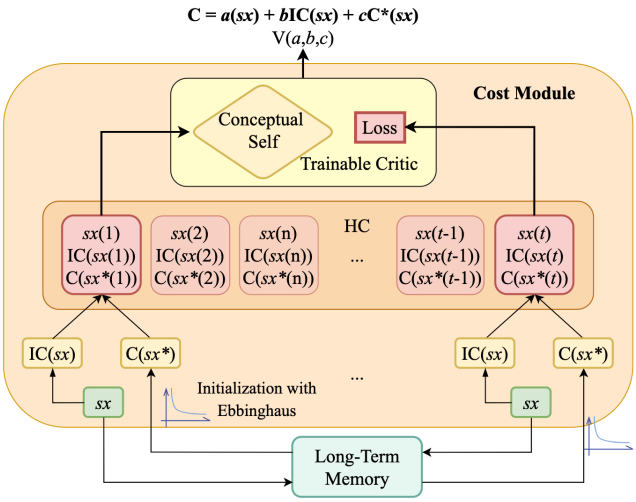

Figure 2: Personalized Cost Module A predefined Intrinsic Cost function (IC) takes the perception of current event(*sx*) or prediction(pred(sx)) in subsequent iteration as its input. Simultaneously, the classifier identifies labels similar to the current *sx* in the LTM categories and their pertinent memory clips(*sx**), returning costs associated with all memory fragments in the same category (C(*sx**)). In the training phase, Conceptual Self retrieves past state vectors HC(*sx*(1)) and the energy of HC(*sx*(*t*)) at a later time. The critic then fine-tunes its parameter vector to minimize the disparity between the target HC(*sx*(1)) and the predicted energy HC(*sx*(*t*)).

Evidently, the consideration of multi-level costs is vital to the personalized and self-consistent nature of human decision-making and makes critical use of the self models. We therefore propose a modification to the cost module to incorporate the Conceptual Self into cost evaluation. In addition to the intrinsic cost representing basic biological discomfort, we introduce a hierarchical system of cost termed higher-level cost (HC). HC considers low-level intrinsic cost while combining personalized information considering the values tied to specific elements in the task situation.

The information pertaining to the states of the world and the self is encoded in variables represented as *sx* and C(*sx**). C(*sx**) denotes the total cost predicted in the preceding event, serving as a baseline for the current cost assessment. This approach is grounded in the assumption that an individual's responses and discomfort-related sentiments tend to exhibit consistency over time. Consequently, it becomes more plausible that the agent would harbor similar feelings toward comparable external stimuli.

Given that C(*sx**) encapsulates the cost from past experiences by the context cues of similar events (similar *sx**), its value is determined by the retrieval process that is similar to the retrieval of past memories in LTM as mentioned in the memory module. The *sx* of current events is juxtaposed with the labels of all categories, returning the cost associated with all discrete task memories in the same category, adjusted with

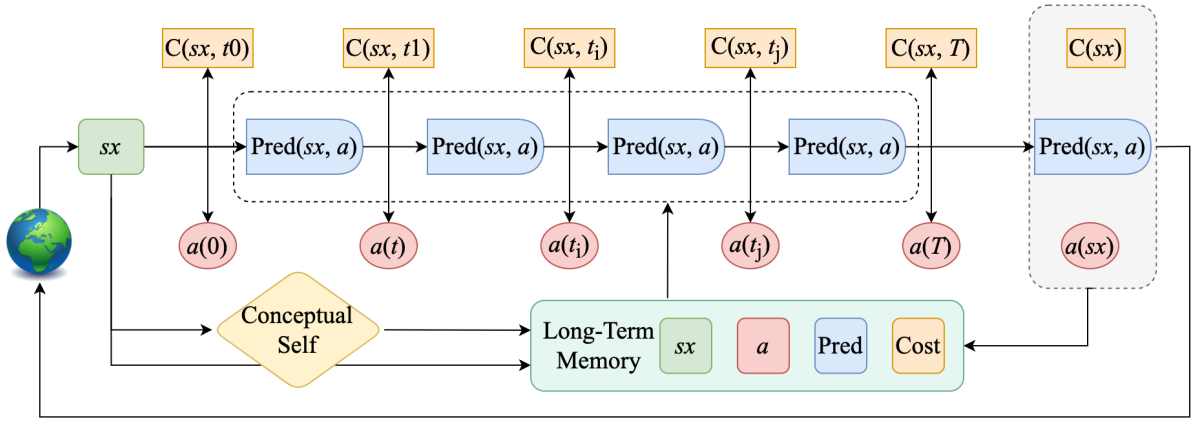

Figure 3: This diagram illustrates the model's Markov process. The perception module estimates the world's state (s0), and the actor proposes a series of actions (stored in $a(0)$). $sx(0)$ also undergoes classification by the Conceptual Self into a memory category. The Pred in the relevant memories among this category is sent to the world model, and their action series to the actor. The world model recursively predicts the estimated world state sequence based on the previous moment's predictions, the Pred fed by the Long-Term Memory (LTM), and the action series. The cost $C(sx(t))$ calculates energy for each predicted state, and iterative modules compute an optimized action series.

weights according to the forgetting curve. HC is thus expressed as the following:

$$\mathrm{HC}(\mathrm{IC}(sx), sx(t), \mathrm{C}^*(sx), t)$$

The three inputs, IC, sx, and $C(sx^*)$, are assigned respective weights through the vector $V(a, b, c)$ in the High-Level Cost (HC). The final cost is computed as:

$$\mathrm{C} = a(sx) + b\mathrm{IC}(sx) + c\mathrm{C}^*(sx)$$

The assignment of values in vector V is determined during training by the Conceptual Self, assessing the importance of each piece of information. The stored HC (time, state, intrinsic energy, previous costs) in the associative short-term memory are accessible for the Conceptual Self to retrieve. During training, the Conceptual Self retrieves a past state vector $HC(sx(1))$ and an intrinsic energy at a later time $HC(sx(\tau))$, adjusting the parameters a, b, c in vector V to minimize the divergence measure between the target and the predicted energy $HC(sx(1))$.

Discussion

In this paper, we propose a modified version of JEPA that aims to enable consistent self-interest and trait-like characteristics in autonomous intelligent agents with the implementation of human-like self models. The modification has two key components: (1) a memory module replacing the latent variable module that supports graded recall of past experiences; (2) a personalized cost module that supports high-level value judgment. Self-important evaluations underlying both processes are powered by the classifier in the memory module, which scans for matching conceptual information between those presented in the task environment and synthesized autobiographical knowledge. Said implementations are inspired by neurocognitive mechanisms underlying the information-processing of human self models comprising autobiographical memory and conceptual self-knowledge.

Self models could provide multiple advantages to intelligent agents in terms of performance. Compared to solely relying on world models, self models enable agents to make personalized and self-interested decisions in complex, open-ended situations by allowing references to synthesized knowledge of past experiences. Such functionalities are especially relevant to the design of autonomous agents aimed to simulate complex human behaviors, which could empower technical applications ranging from chatbots to immersive social environments. Park et al. (2023) proposed an extended large language models (LLM) architecture featuring a memory system which they demonstrated could power generative agents with social interactions. However, said memory system does not support human-like processing of autobiographical knowledge, but rather records all past experiences without decay and uses a universal criterion for importance evaluation during retrieval, which could lead to difficulties in long-term goal-directed planning and believability in complex social situations. In that respect, we suggest that our architecture may offer an alternative path that could lead to better self-consistency in similar applications with the implementation of self models.

As this is a position paper, there are several aspects of the present architecture that await specifications. For instance, feedback administration is needed to prevent cases where the model is self-consistent but does not align with external inputs and outputs (i.e. delusional), which could happen to the current instantiation given it only relies on predicted outcome (sy) for self-supervised learning. Overall, the proposal outlined in this paper offers tractable future directions toward building cognitive-inspired machine agents that could reason and act in complex open-ended situations with human-level self-consistency.

Acknowledgements

The authors would like to thank Stan Klein, Benjamin Liu, Zhaoting Liu, Peiyang Song as well as the anonymous reviewers for their valuable insights and comments.

References

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N. (2023). Self-supervised learning from images with a joint-embedding predictive architecture. *arXiv preprint arXiv:2301.08243*.

Brown, T. I., Rissman, J., Chow, T. E., Uncapher, M. R., & Wagner, A. D. (2018). Differential medial temporal lobe and parietal cortical contributions to real-world autobiographical episodic and autobiographical Semantic Memory. *Scientific Reports*, *8*(1).

Cassenti, D. N., Veksler, V. D., & Ritter, F. E. (2022). Editor's review and introduction: Cognition-inspired Artificial Intelligence. *Topics in Cognitive Science*, *14*(4), 652–664.

Conway, M. A. (2005). Memory and the self. *Journal of Memory and Language*, *53*(4), 594–628.

Conway, M. A., & Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self-memory system. *Psychological Review*, *107*(2), 261–288.

Conway, M. A., Singer, J. A., & Tagini, A. (2004). The self and autobiographical memory: Correspondence and coherence. *Social Cognition*, *22*(5), 491–529.

Davis, R. L., & Zhong, Y. (2017). The biology of forgetting—a perspective. *Neuron*, *95*(3), 490–503.

Demiray, B., & Bluck, S. (2011). The relation of the conceptual self to recent and distant autobiographical memories. *Memory*, *19*(8), 975–992.

D'Argembeau, A. (2013). On the role of the ventromedial prefrontal cortex in self-processing: The valuation hypothesis. *Frontiers in Human Neuroscience*, *7*.

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology*. Dover.

Eslami, S. M., Jimenez Rezende, D., Besse, F., Viola, F., Morcos, A. S., Garnelo, M., Ruderman, A., Rusu, A. A., Danihelka, I., Gregor, K., Reichert, D. P., Buesing, L., Weber, T., Vinyals, O., Rosenbaum, D., Rabinowitz, N., King, H., Hillier, C., Botvinick, M., Wierstra, D., Kavukcuoglu, K., Hassabis, D. (2018). Neural scene representation and rendering. *Science*, *360*(6394), 1204–1210.

Friston, K., Moran, R. J., Nagai, Y., Taniguchi, T., Gomi, H., & Tenenbaum, J. (2021). World model learning and inference. *Neural Networks*, *144*, 573–590.

Grilli, M. D., & Verfaellie, M. (2015). Supporting the self-concept with memory: Insight from amnesia. *Social Cognitive and Affective Neuroscience*, *10*(12), 1684–1692.

Ha, D., & Schmidhuber, J. (2018). World models. *arXiv preprint arXiv:1803.10122*.

Haber, N., Mrowca, D., Li, F., Yamins, D. L. K. (2018). Learning to play with intrinsically-motivated, self-aware agents. *arXiv preprint arXiv:1802.07442*.

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

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *The Journal of Neuroscience*, *26*(32), 8360–8367.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired Artificial Intelligence. *Neuron*, *95*(2), 245–258.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Krauss, P., & Maier, A. (2020). Will We Ever Have Conscious Machines?. *Frontiers in computational neuroscience*, *14*, 556544.

Kenrick, D. T., Griskevicius, V., Neuberg, S. L., & Schaller, M. (2010). Renovating the pyramid of needs. *Perspectives on Psychological Science*, *5*(3), 292–314.

Kihlstrom, J. F., & Cantor, N. (1984). Mental representations of the self. *Advances in Experimental Social Psychology*, 1–47.

Kihlstrom, J. F., Albright, J. S., Klein, S. B., Cantor, N., Chew, B. R., & Niedenthal, P. M. (1988). Information processing and the study of the self. *Advances in Experimental Social Psychology*, 145–178.

Klein, S. B., & Loftus, J. (1993). The mental representation of trait and autobiographical knowledge about the self. In T. K. Srull & R. S. Wyer, Jr. (Eds.), *The mental representation of trait and autobiographical knowledge about the self*. Lawrence Erlbaum Associates, Inc.

LeCun, &. (2022). *A path towards autonomous machine intelligence, Version 0.9.2, 2022-06-27*.

Levorsen, M., Aoki, R., Matsumoto, K., Sedikides, C., & Izuma, K. (2023). The self-concept is represented in the medial prefrontal cortex in terms of self-importance. *The Journal of Neuroscience*, *43*(20), 3675-3686.

Matthews, G., Deary, I. J., & Whiteman, M. C. (2003). *Personality Traits*. Cambridge University Press.

Moore, D. A., & Loewenstein, G. (2004). Self-interest, automaticity, and the psychology of conflict of interest. *Social Justice Research*, *17*(2), 189–202.

Neisser, U. (1988). Five kinds of self-knowledge. *Philosophical Psychology*, *1*(1), 35–59.

Northoff, G. (2013). Brain and self – a neurophilosophical account. *Child and Adolescent Psychiatry and Mental Health*, *7*(1), 28.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. *arXiv preprint arXiv:2304.03442*.

Pelau, C., Dabija, D.-C., & Ene, I. (2021). What makes an AI device human-like? the role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior*, *122*, 106855.

Prebble, S. C., Addis, D. R., & Tippett, L. J. (2013). Autobiographical memory and sense of self. *Psychological Bulletin*, *139*(4), 815–840.

Rao, R. P., Gklezakos, D. C., & Sathish, V. (2023). Active predictive coding: A unifying neural model for active

perception, compositional learning, and hierarchical planning. *Neural Computation*, *36*(1), 1–32.
Schiller, D., & Phelps, E. A. (2011). Does reconsolidation occur in humans? *Frontiers in Behavioral Neuroscience*, *5*.
Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T., & Silver, D. (2020). Mastering Atari, go, chess and shogi by planning with a learned model. *Nature*, *588*(7839), 604–609.
Sheldon, S., Fenerci, C., & Gurguryan, L. (2019). A neurocognitive perspective on the forms and functions of autobiographical memory retrieval. *Frontiers in Systems Neuroscience*, *13*.
Sheldon, S., McAndrews, M. P., & Moscovitch, M. (2011). Episodic memory processes mediated by the medial temporal lobes contribute to open-ended problem solving. *Neuropsychologia*, *49*(9), 2439–2447.
Simon, H. A., Dantzig, G. B., Hogarth, R., Plott, C. R., Raiffa, H., Schelling, T. C., Shepsle, K. A., Thaler, R., Tversky, A., & Winter, S. (1987). Decision making and problem solving. *Interfaces*, *17*(5), 11–31.
Summerfield, C. (2022). *Natural general intelligence: How understanding the brain can help us build AI*. Oxford university press.
Talamini, L. M., & Gorree, E. (2012). Aging memories: differential decay of episodic memory components. *Learning & Memory*, *19*(6), 239–246.
Tulving, E. (1984). Relations among components and processes of memory. *Behavioral and Brain Sciences*, *7*(2), 257–268.
Vaidya, A. R., & Badre, D. (2020). Neural systems for memory-based value judgment and decision-making. *Journal of Cognitive Neuroscience*, *32*(10), 1896–1923.
Vogeley, K., Kurthen, M., Falkai, P., & Maier, W. (1999). Essential functions of the human self model are implemented in the prefrontal cortex. *Consciousness and Cognition*, *8*(3), 343–363.