frontiers | Frontiers in Computational Neuroscience

TYPE Original Research
PUBLISHED 20 February 2023
DOI 10.3389/fncom.2023.1099593

Check for updates

OPEN ACCESS

EDITED BY
Mario Senden,
Maastricht University, Netherlands

REVIEWED BY
Thomas Parr,
University College London, United Kingdom
Francisco Barceló,
University of the Balearic Islands, Spain

*CORRESPONDENCE
Zhizhuo Yang
✉ zy8981@rit.edu

RECEIVED 16 November 2022
ACCEPTED 30 January 2023
PUBLISHED 20 February 2023

CITATION
Yang Z, Diaz GJ, Fajen BR, Bailey R and Ororbia AG (2023) A neural active inference model of perceptual-motor learning. *Front. Comput. Neurosci.* 17:1099593. doi: 10.3389/fncom.2023.1099593

A neural active inference model of perceptual-motor learning

Zhizhuo Yang[1]*, Gabriel J. Diaz[2], Brett R. Fajen[3], Reynold Bailey[1] and Alexander G. Ororbia[1]

[1]Golisano College of Computing and Information Sciences, Rochester Institute of Technology, Rochester, NY, United States, [2]Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology, Rochester, NY, United States, [3]Department of Cognitive Science, Rensselaer Polytechnic Institute, Troy, NY, United States

The active inference framework (AIF) is a promising new computational framework grounded in contemporary neuroscience that can produce human-like behavior through reward-based learning. In this study, we test the ability for the AIF to capture the role of anticipation in the visual guidance of action in humans through the systematic investigation of a visual-motor task that has been well-explored—that of intercepting a target moving over a ground plane. Previous research demonstrated that humans performing this task resorted to anticipatory changes in speed intended to compensate for semi-predictable changes in target speed later in the approach. To capture this behavior, our proposed "neural" AIF agent uses artificial neural networks to select actions on the basis of a very short term prediction of the information about the task environment that these actions would reveal along with a long-term estimate of the resulting cumulative expected free energy. Systematic variation revealed that anticipatory behavior emerged only when required by limitations on the agent's movement capabilities, and only when the agent was able to estimate accumulated free energy over sufficiently long durations into the future. In addition, we present a novel formulation of the *prior mapping function* that maps a multi-dimensional world-state to a uni-dimensional distribution of free-energy/reward. Together, these results demonstrate the use of AIF as a plausible model of anticipatory visually guided behavior in humans.

KEYWORDS

interception, locomotion, active inference, learning, anticipation

1. Introduction

The active inference framework (AIF) (Friston et al., 2009) is an emerging theory of neural encoding and processing that captures a wide range of cognitive, perceptual, and motor phenomena, while also offering a neurobiologically plausible means of conducting reward-based learning through the capacity to predict sensory information. The behavior of an AIF agent involves the selection of action-plans that span into the near future and centers around the learning of a probabilistic generative model of the world through interaction with the environment. Ultimately, the agent must take action such that it is making progress toward its goals (goal-seeking behavior) while also balancing the drive to explore and understand its environment (information maximizing behavior), adjusting the internal states of its world to better account for the evidence that it acquires over time. As a result, AIF unifies perception, action, and learning by framing them as processes that result from approximate Bayesian inference.

The AIF framework has been used to study a variety of reinforcement learning (RL) tasks, including the inverted pendulum problem (*CartPole*) (Millidge, 2020; Shin et al., 2022), the mountain car problem (*MountainCar*) (Friston et al., 2009; Ueltzhöffer, 2018; Çatal et al., 2020; Tschantz et al., 2020a; Shin et al., 2022) and the frozen lake problem (*Frozen Lake*)

(Sajid et al., 2021). Each task places different demands on motor and cognitive abilities. For instance, *CartPole* requires online control of a paddle to balance a pole upright, whereas *MountainCar* requires intelligent exploration of the task environment; a simple "greedy" policy (typical of many modern-day RL approaches) would fail to solve the problem. The popular *Frozen Lake* requires skills related to spatial navigation and planning if the agent is to find the goal while avoiding unsafe states.

One fundamental aspect of human and animal behavior that has so far not been sufficiently studied from an active inference perspective is the on-line visual guidance of locomotion. On-line visual guidance comprises a class of ecologically important behaviors for which movements of the body are continuously regulated based on currently available visual information seen from the first-person perspective. Some of the most extensively studied tasks include steering toward a goal (Warren et al., 2010), negotiating complex terrain on foot (Matthis and Fajen, 2013; Diaz et al., 2018), intercepting moving targets (Fajen and Warren, 2007), braking to avoid a collision (Yilmaz and Warren, 1995; Fajen and Devaney, 2006), and intercepting a fly ball (Chapman, 1968; Fajen et al., 2008). For each of these tasks, researchers have formulated control strategies that capture the coupling of visual information and action.

One aspect of on-line visual guidance that AIF might be particularly well-suited to capture is anticipation. To successfully perform any of these kinds of tasks, actors must be able to regulate their actions in anticipation of future events. One approach to capturing anticipation in visual guidance is to identify sources of visual information that specify how the actor should move at the current instant in order to reach the goal in the future. For example, when running to intercept a moving target, the sufficiency of the interceptor's current speed is specified by the rate of change in the exocentric visual direction of the target, or *bearing angle* (Figure 1). If the interceptor is able to move so as to maintain a constant bearing angle (CBA), then an interception is guaranteed. Such accounts of anticipation are appealing because they avoid the need for planning on the basis of predictions or extrapolations of the agent's or target's motion, thereby presumably requiring fewer cognitive resources for task execution. Similar accounts of anticipation in the context of locomotor control have been developed for fly ball catching (Chapman, 1968) and braking (Lee, 1976).

However, there are other aspects of anticipatory control that are more difficult to capture based on currently available information alone. For example, moving targets sometimes change speeds and directions in ways that are somewhat predictable, allowing actors to alter their movement in advance in anticipation of the most likely change in target motion. This was demonstrated in a previously published study in which subjects were instructed to adjust their self-motion speed while moving along a linear path in order to intercept a moving target that changed speed partway through each episode (Diaz et al., 2009). Note that episode refers to a single, complete course of interception for the agent and the target to be compatible with the conventions used by the reinforcement learning community. In contrast, Diaz et al. (2009) uses the word *trial*. The final target speed randomly varied between episodes such that the target usually accelerated but occasionally decelerated. In response, subjects quickly learned to adjust their speed during the first part of the episode in anticipation of the change in target speed that was most likely given past experience and the initial conditions of that episode.

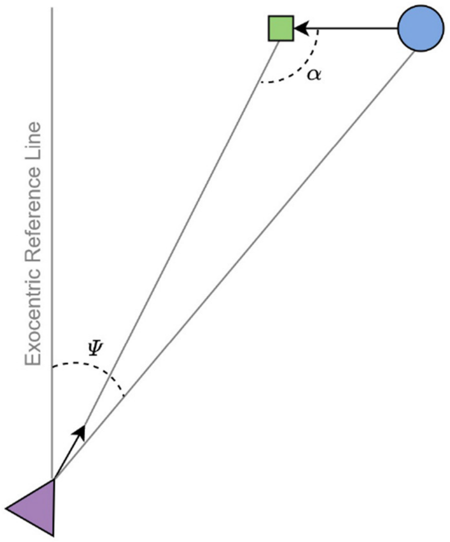

FIGURE 1
A top-down view of the interception problem. The agent (triangle) and target (circle) approach the invisible interception point (square) by going straight ahead. Ψ denotes the exocentric direction of the target (bearing angle) and α denotes the target's approach angle. Image adapted from Diaz et al. (2009).

Active inference offers a potentially useful framework for understanding and modeling this kind of anticipatory behavior. The behavior of an AIF agent involves the selection of action plans (or policies) that span into the near future. These plans are selected based on *expected free energy* (EFE), i.e., a reward signal that takes into account both the action's contribution to reaching a desired goal state (i.e., an *instrumental* component), and the new information gained by the action (i.e., an *epistemic* component). This method of action selection is ideal for the study of predictive and anticipatory behavior in that it allows for the selection of action plans that do not immediately contribute to task completion, but that reveal to the agent something previously unknown about how the agent's action affects the environment. Similarly, in the task presented in Diaz et al. (2009), the human participants learned that success required increasing speed early in the episode in order to increase the likelihood of an interception after the target's semi-predictable change in speed. Critically, this early change in speed was not motivated by currently available visual information, but rather by the positive reinforcement of actions selected in the process of task exploration.

In contrast to reinforcement learning methods, active inference (AIF) formulates action-driven learning and inference from a Bayesian, belief-based perspective (Parr and Friston, 2019; Sajid et al., 2021). Generally, AIF offers: (1) flexibility to define a prior preference (or preferred outcome) over the observation space (which pushes the agent to uncover goal-orienting policies), which provides an alternative to designing a reward function, (2) a principled treatment for epistemic exploration as a means of uncertainty reduction, information gain, and intrinsic motivation (Parr and Friston, 2017, 2019; Schwartenbeck et al., 2019), and (3) an encompassing uncertainty or precision over the beliefs that the generative model of the AIF agent computes as a natural part of then agent's belief updating (Parr and Friston, 2017). Despite being a popular and powerful framework of perception, action (Friston, 2009, 2010; Buckley et al., 2017; Friston K. et al., 2017), decision-making and planning (Kaplan and Friston, 2018; Parr and Friston, 2018) with biological plausibility, AIF has been mostly applied to problems with a low-dimensionality and often discrete state space and actions (Friston et al., 2012, 2015, 2018; Friston K. et al., 2017; Friston K. J. et al., 2017). One of the key limitations is that calculation of the EFE values for all policies starting from the current time step is needed in order to select the optimal action at the immediate time step. The exact EFE calculation becomes intractable quickly as the size of the action space $|\mathcal{A}|$ and the planning time horizon H grows (Millidge, 2020; Shin et al., 2022). We refer to Da Costa et al. (2020) for a comprehensive review on AIF.

The present study makes several specific contributions to the understanding of visually guided action and active inference:

- We present a novel model for locomotor interception of a target that changes speeds semi-predictably, as in Diaz et al. (2009). This model is a scaled-up version of AIF where EFE is treated as a negative value function in reinforcement learning (RL) (Shin et al., 2022) and deep RL methodology is utilized to scale AIF to solve tasks such as locomotor interception with continuous state spaces. Specifically, our method predicts action-conditioned EFE values with a *joint* network (see Section 2.4.2) and by bootstrapping on the continuous observation space over a long time horizon. This allows the agent to account for the long-term effects of its current chosen action(s).
- To calculate the *instrumental* value, we designed a problem-specific *prior mapping function* to convert the original observations into a one-dimensional prior space where a prior preference can be (more easily) specified. This allows us to inject domain knowledge into the *instrumental* reward. The *instrumental* measurements in prior space simultaneously promote interpretability as well as computationally efficient task performance.
- We present a comparison of task performance of a baseline deep-Q network (DQN) agent, or an AIF agent in which EFE is computed using only the *instrumental* signal/component, with a full AIF agent in which EFE is computed using both *instrumental* and *epistemic* signals/components.
- We demonstrate behavioral differences among our full AIF agent under the influence of two varying parameters: the discount factor γ, which describes the weight on future accumulated quantities when calculating EFE value at each time step, and pedal lag coefficient K, which specifies how responsive changes in pedal position is reflected on agent's speed (or the amount of inertia that is associated with the agent's vehicle).
- We interpret our findings as a model for anticipation in the context of visually guided action as well as in terms of specific contributions to the active inference and machine learning communities.

2. Materials and methods

Our aim in this study was to develop an agent that selects from a set of discrete actions in order to perform the task of interception. In this section, we describe the task that we aim to solve as well as formally describe the AIF model designed to tackle it. We start with the problem formulation and brief notation and definitions, then move on to describe our proposed agent's inference and learning dynamics.

2.1. The perceptual-motor problem: Intercepting a moving target

We designed and simulated a perception-motor problem based on the human interception task used by Diaz et al. (2009). In the original study, subjects sat in front of a large rear-projection screen depicting an open field with a heavily textured ground plane. The subject's task was to intercept a moving spherical target by controlling the speed of self-movement along a linear trajectory with a foot pedal, the position of which was mapped onto speed according to a first-order lag. Subjects began each episode from a stationary position at an initial distance sampled uniformly from between 25 and 30 m from the interception point. The spherical target approached the subject's path at one of three initial speeds (11.25, 9.47, and 8.18 m/s). Between 2.5 and 3.25 s after the episode began, the target changed speeds linearly by an amount that was sampled from a normal distribution of possible final speeds. The mean of the distribution was 15 m/s such that target speed usually increased, but occasionally decreased (standard deviation was 5 m/s, final speed is truncated by one standard deviation from the mean). The change of target speed takes exactly 500 ms.

This interception problem is difficult because a human or agent that is purely reactive to the likely change in target speed will often arrive at the interception point after the target (e.g., they will be too slow). The problem is exacerbated when the agent's vehicle is less responsive. In Diaz et al. (2009), subjects were found to increase their speed during the early part of the episode in order to anticipate the most likely change in target speed, which helped them perform at near optimal levels. Differences between the behavior of subjects and the ideal pursuer were also found under some conditions. Findings in the original study further yielded insight into the strategies that humans adopt when dealing with uncertainty in realistic interception tasks.

2.2. Notation

We next define the notation and mathematical operators that we will use throughout the rest of this paper. $\odot$ indicates a Hadamard product, $\cdot$ indicates a matrix/vector multiplication (or dot product if the two objects it is applied to are vectors of the same shape), and $(\mathbf{v})^T$ denotes the transpose. $||\mathbf{v}||_p$ is used to represent the p-norm where $p = 2$ results in the 2-norm or Euclidean (L2) distance.

2.3. Action and input space specification

To simplify the problem for this work, we assume that the mapping between environmental (latent) states and observations is the identity matrix. Furthermore, we formulate the problem as a Markov Decision Process (MDP) with a discrete action space. The action space $\mathbf{a}_t$ (action vector at time t) is defined as a one-hot vector $\mathbf{a} \in \{0,1\}^{6\times 1}$, where each dimension corresponds to a unique action and the actions are mutually exclusive. Each dimension corresponds to one of the pedal speeds (m/s) in $\{2, 4, 8, 10, 12, 14\}$ respectively. Once a pedal speed is selected, the agent will change its own speed by the amount of $\Delta V = K{*}(V_p - V_s)$ in one time step where V_p is pedal speed, V_s is current subject speed and K is a constant lag coefficient. In this study, we experiment with 2 variants of peal lag coefficient, i.e., $K = 1.0K'$ and $K = 0.5K'$. K' is set to 0.017 to be consistent with the original study (Diaz et al., 2009) and provides a smooth relationship between the pedal movement and vehicle speed change. Similar to Tschantz et al. (2020b), we assume that the control state vector (which, in AIF, control states are originally treated separately from action states) lines up one-to-one with the action vector, meaning that it too is a vector of the form $\mathbf{u} \in \{0,1\}^{6\times 1}$. We define the observation/state space ($\mathbf{o} \in \mathcal{R}^{4\times 1}$) to be a 4-dimensional vector $\mathbf{o}_t = \langle x_t, v_t, x_s, v_s\rangle^T$, which corresponds to target distance, target speed, subject distance and subject speed. All distances aforementioned are with respect to the invisible interception point.

2.4. Neural active inference

Active inference (AIF) is a Bayesian computational framework that brings together perception and action under one single imperative: minimizing *free energy*. It accounts for how self-organizing agents operate in dynamic, non-stationary environments (Friston, 2019), offering an alternative to standard, reward function-centric reinforcement learning (RL). In this study, we craft a simple AIF agent that resembles Q-learning (Shin et al., 2022) where the *expected free energy* (EFE) serves the role of a negative action-value function in RL. We frame the definition of EFE in the context of a stochastic policy and cast action-conditioned EFE as a negative action-value using a policy $\phi = \phi(a_t|\mathbf{s}_t)$ (where $\mathbf{s}_t = \mathbf{o}_t$ as per our assumption earlier). The same policy ϕ is used for each future time step τ, and the probability distribution over the first-step action is separated from ϕ resulting in a substitution distribution $q(a_t)$ for $\phi(a_t)$. Therefore, the one-step substituted EFE can be interpreted as the EFE of a policy of $(q(a_t), \phi(a_{t+1}), \ldots, \phi(a_T)))$.

Following Shin et al. (2022), we consider the deterministic optimal policy ϕ^* which always seeks an action with a minimum EFE and obtain the following EFE definition:

$$\begin{aligned} G_{\phi^*}(s_t) &= \min_a G_{\phi^*}(s_t|a) \\ &= \min_a \mathbb{E}_{p(s_{t+1}|s_t,a_t=a)p(o_{t+1}|s_{t+1})} \\ &\left[log\frac{p(s_{t+1}|s_t, a_t=a)}{\tilde{p}(o_{t+1})q(s_{t+1}|o_{t+1})} + G_{\phi^*}(s_{t+1}) \right] \end{aligned} \quad (1)$$

According to Shin et al. (2022), the equation above is quite similar to the Bellman optimality equation, where $G_{\phi^*}(s_{t+1})$ corresponds to the state-value function $V^*(s_{t+1}) = \max_\pi V^\pi(s_{t+1})$ and $G_{\phi^*}(s_t|a)$ corresponds to the action-value function $Q^*(s_t, a)$. Then the first term $log\frac{p(s_{t+1}|s_t,a_t=a)}{\tilde{p}(o_{t+1})q(s_{t+1}|o_{t+1})}$ can be treated as a one-step negative reward and thus EFE can be treated as a negative value function. This term can then be further decomposed into two components:

$$\begin{aligned} R_t := -log\frac{p(s_{t+1}|s_t, a_t=a)}{\tilde{p}(o_{t+1})q(s_{t+1}|o_{t+1})} &= \underbrace{log\tilde{p}(o_{t+1})}_{Instrumental} + \underbrace{(-log\frac{p(s_{t+1}|s_t, a_t)}{q(s_{t+1}|o_{t+1})})}_{Epistemic} \\ &= R_{t,i} + R_{t,e} \end{aligned} \quad (2)$$

To connect the formulation above back to AIF, with the term rephrased as $R_{t,i}$ is the instrumental (also known as *extrinsic*, *pragmatic* or *goal-directed*) component (Tschantz et al., 2020b), which measures the similarity between the future outcome following the policy ϕ and preferred outcome (or prior preference). The term rephrased as $R_{t,e}$ is known as the epistemic (also known as *intrinsic*, *uncertainty-reducing* or *information-seeking*) component (Tschantz et al., 2020b), which measures the prediction error between the estimation of the future state by the transition model and the state predicted by the encoder given the actual observation from the environment.

Ultimately, we simplify and approximate the search for optimal EFE values by adapting an estimation approach based on the Bellman equation, arriving at a Q-learning bootstrap scheme. We assume that the outcome/observation can be set equal to state variables and, as a result, our generative model is designed with respect to fully observed environment (Tschantz et al., 2020a). Following the active inference literature, we adopt the Laplace assumption and mean-field approximation. Therefore, a fixed identity covariance matrix is used for the likelihood distribution $p(o|s)$. Our model (the function approximator) outputs the mean of states, which encodes the belief that there is a direct mapping between outcomes and states. Similarly, our model outputs the mean of estimated EFE values. Following (Mnih et al., 2015), we integrated an experience replay as well as a target network in order to facilitate learning and improve sample efficiency. Note that the Q-learning style framing of negative EFE estimation is referred to as G-learning. Our model estimates the EFE for each possible action that it could take in the immediate next time step (i.e., time $t + 1$) then selects the action that corresponds to the minimal EFE value. This, in effect, corresponds to only explicitly calculating the EFE over a horizon of 1 (whereas as planning over horizons >1 quickly become prohibitive, requiring expensive search methods such as Monte Carlo tree search) but incorporates a bootstrap estimate of future EFE values via the G-learning setup. Our definition in Equation (1) is similar to the EFE definition in Friston et al. (2021) in the sense that EFE is formulated recursively in both works. However, differences between our method and sophisticated inference (Friston et al., 2021) still exist. For instance, our method works with continuous state space whereas sophisticated inference works with a discrete state space. Our method displays a connection to Q-learning, thus it is able to plan over a trajectory of arbitrary length in principle using bootstrap estimation, whereas sophisticated inference terminates the evaluation of recursive EFE whenever an action is found as unlikely or an outcome is implausible. We utilize the AIF framework within the G-learning framing for the interception task and modify the framework to fit the interception task, see Figure 2. Spatial variables, i.e., distance and speed, will serve as the inputs to our framework and, as mentioned before, an identity mapping is assumed to connect the observation directly to the state variables (allowing us to avoid having to learn additional parameterized encoder/decoder functions). As

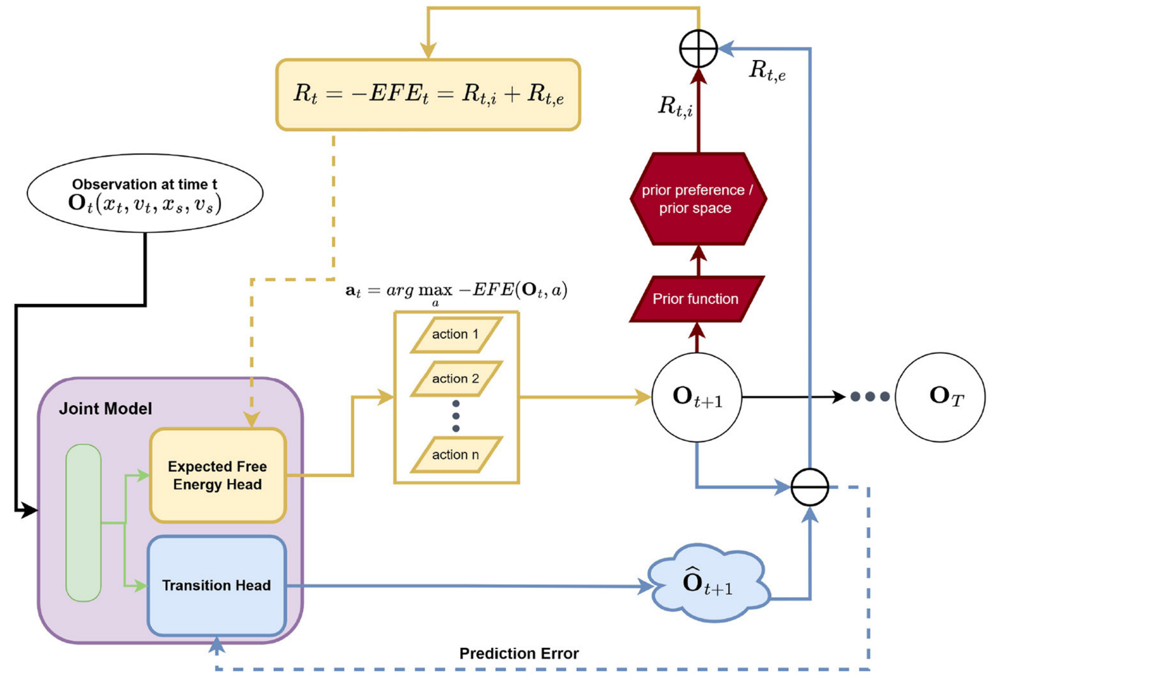

FIGURE 2
Our neural AIF architecture for the interception task. The joint model is a two-headed artificial neural network which consists of shared hidden layers, an EFE (estimation) head, and a transition dynamics (prediction) head. The EFE head estimates EFE values for all possible actions given the current (latent/hidden) state. An action which is associated with maximum EFE value is selected and executed in the environment and the resulting observation is fed into the *prior mapping function* where the *instrumental* value $R_{t,i}$ is calculated in prior space. Meanwhile, the transition dynamics head predicts the resulting observation given the current (latent/hidden) state. The error between the predicted and actual observation at $t+1$ forms the *epistemic* value $R_{t,e}$. The summation of $R_{t,i}$ and $R_{t,e}$ results in the final EFE (target) value.

a result, the AIF agent we designed for this paper's experiments consists of two major components: a *prior mapping function* and a multi-headed joint neural model.

Notably, our particular proposed joint model works jointly as a function approximator of EFE values as well as a forward dynamics predictor. It takes in the current observation $\mathbf{o}_t$ as input and then conducts, jointly, action selection and next-state prediction (as well as *epistemic* value estimation). The selected action is executed and the resulting observation is returned by the environment. The *prior mapping function* itself takes in as input the next observation $\mathbf{o}_{t+1}$, the consequence/result of the agent's currently selected action, and calculates the log likelihood of the preferred/prior distribution (set according to expert knowledge related to the problem), or the *instrumental* term $R_{t,i}$. The squared difference between the outcome of the selected action $\mathbf{o}_{t+1}$ and its estimation $\hat{\mathbf{o}}_{t+1}$ (as per the generative transition component of our model) forms the *epistemic* term $R_{t,e}$ as shown in Equation 7. The summation of the *instrumental* and *epistemic* terms forms the G-value (or negative EFE value) which is ultimately used to train / adapt the joint model. Formally, $R_t = -G_{\phi^*}(s_t) = R_{t,i} + R_{t,e}$. We explain each component in detail below.

2.4.1. The prior mapping function and prior space

With the ability and freedom of designing a prior preference (or distribution over problem goal states or preferred outcomes) afforded by AIF, we integrate domain knowledge of the interception task into the design of a *prior mapping function*. In essence, our designed *prior mapping function* transforms the original observation vector $\mathbf{o}_t$ to a lower-dimensional space (the prior space) where a semantically meaningful variable is calculated and prior preference distribution is specified over this new variable—in our case, this is set to be the *speed difference*, as shown in Figure 3. The *speed difference* represents the difference between the agent's speed after taking the selected action and the speed required for successful interception, i.e., *speed difference* $=$ *speed*$_{agent}$ $-$ *speed*$_{required}$. Given the current observation, the required speed is calculated as the agent's distance to the interception point divided by the first-order target time-to-contact (TTC). We define TTC as the duration for the target or agent to reach the theoretical interception point from the current time step regardless of the success of the actual interception task. Then, target's first-order TTC is the amount of time that it would take for the target to reach the interception point assuming that target speed does not change throughout the episode, i.e., $TTC_{first-order} = x_t/v_{constant}$ where x_t is the target distance and $v_{constant}$ is the target constant speed. In our neural AIF framework, the *instrumental* values are calculated given all possible actions (blue circles in Figure 3) and a prior distribution over *speed difference*. The smaller the absolute *speed difference* associated with a particular action, the higher the *instrumental* value *prior mapping function* assigns.

Note that the agent might not have enough time to adjust its speed later in the interception task if it only follows the guidance of this *prior mapping function* without anticipating the likely future speed change of target, since this *prior mapping function* only accounts/embodies first-order information. To overcome this limitation, we investigated the effects of discounted long-term EFE value on the behavior of the agent in Section 3.4.

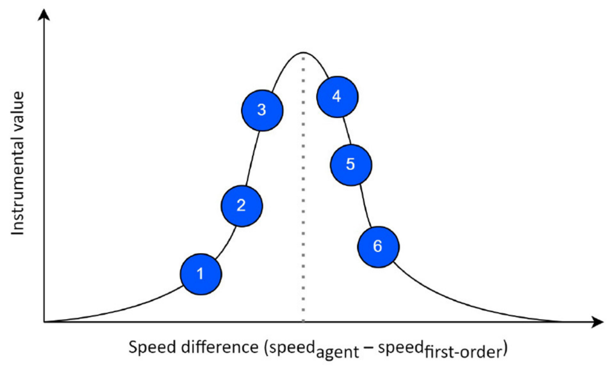

FIGURE 3
The prior preference specified in the prior space where each action corresponds to a different *instrumental* value. Circles correspond to pedal positions to choose from.

2.4.2. Joint model

Our proposed joint model embodies two key functionalities: EFE estimation and transition dynamics prediction, which are typically implemented as separate artificial neural networks (ANNs) in earlier AIF studies (Shin et al., 2022) (in contrast, we found that, during preliminary experimentation, that a joint, fused architecture improved both the agent's overall generalization ability as well as its training stability). Concretely, we implement the joint model as a multi-headed ANN with an EFE head and a transition head (see Figure 2). The system takes in the current observation $\mathbf{o}_t$ and predicts: (1) the EFE values for all possible actions, and (2) a future observation at a distance $\mathbf{o}_{t+D}$ (in this work, we fix the temporal distance to be one step, i.e., $D = 1$). Within the joint model, the current observation $\mathbf{o}_t$ is taken as input and a latent hidden activity vector $\mathbf{z}_t$ is produced, which is then provided to both output heads as input. The transition head $p(\mathbf{o}_{t+D}|\mathbf{z}_t)$ serves as a generative model (or a forward dynamics model) and the EFE head $G_{\phi^*}(\mathbf{z}_t, a)$ represents an approximation over the EFE values. As a result, EFE module and transition modules are wired together such that the prediction of the future observation $\mathbf{o}_{t+D}$ and the estimation of EFE values $\mathbf{G}_{t+D}$ are driven by the shared encoding from the topmost (hidden) layer of the joint model. This enables the sharing of underlying knowledge between the module selecting actions and the module predicting the outcome(s) of an action. Our intuition is that we humans tend to evaluate the "value" of an action by the consequences that it produces.

We next formally describe the dynamics of our joint model, including both its inference and learning processes.

Inference In general, our agent is meant to produce an action conditioned on observations (or states) sampled from the environment at particular time-steps. Specifically, within any given T-step episode, our agent receives as input the observation $\mathbf{o}_t \in \mathcal{R}^{D\times 1}$, where D is the dimensionality of the observation space $\mathbf{o}_t$ ($D =$ 4 for the problem investigated in this study). The agent then produces a set of approximate free energy values, one for each action (similar in spirit to Q-values) as well as a prediction of the next observation that it is to receive from its environment (i.e., the perceptual consequence of its selected action).

Formally, in this work, the outputs described above are ultimately produced by a multi-output function $\mathbf{z}^3_a, \mathbf{z}^3_o = f_\Theta(\mathbf{o}_t)$, implemented as a multi-layer perceptron (MLP), where $\mathbf{z}^3_a$ contains estimated expected free energy values (one per discrete action) while $\mathbf{z}^3_o$ is the generative component's estimation of the next incoming observation $\mathbf{o}_{t+1}$. Note that we denote only outputting an action value set from this model as $\mathbf{z}^3_a = f^a_\Theta(\mathbf{o}_t)$ (using only the action output head) and only outputting an observation prediction as $\mathbf{z}^3_o = f^o_\Theta(\mathbf{o}_t)$ (using only the state prediction head). This MLP is parameterized by a set of synaptic weight matrices $\Theta = \{\mathbf{W}^1, \mathbf{W}^2, \mathbf{W}^3_a, \mathbf{W}^3_o\}$, that operates according to the following:

$$\mathbf{z}^1 = \phi_z(\mathbf{W}^1 \cdot \mathbf{z^0}),\ \mathbf{z}^2 = \phi_z(\mathbf{W}^2 \cdot \mathbf{z^1}) \tag{3}$$

$$\mathbf{z}^3_a = \phi_a(\mathbf{W}^3_a \cdot \mathbf{z}^2)),\ \mathbf{z}^3_o = \phi_o(\mathbf{W}^3_o \cdot \mathbf{z}^2)) \tag{4}$$

Where $\mathbf{z}^0 = \mathbf{o}_t$ (the input layer to our model is the observation at t). Note that a single discrete action is read out/chosen from our agent function's action output head as: $a = \arg\max_a f^a_\Theta(\mathbf{o}_t)$. The linear rectifier $\phi_z(\mathbf{v}) = \max(0, \mathbf{v})$ was chosen to be the activation function applied to the internal layers of our model while $\phi_a(\mathbf{v}) = \mathbf{v}$ (the identity) is the function specifically applied to the action neural activity layer $\mathbf{z}^3_a$ and $\phi_o(\mathbf{v}) = \mathbf{v}$ is the function applied to predicted observation layer neurons. Note that the first hidden layer $\mathbf{z}^1 \in \mathcal{R}^{J_1\times 1}$ contains J_1 neurons and $\mathbf{z}^2 \in \mathcal{R}^{J_2\times 1}$ contains J_2 neurons, respectively. The action output layer $\mathbf{z}^3_a \in \mathcal{R}^{A\times 1}$ contains A neurons ($A = 6$ for the problem investigated in this study), one neuron per discrete action (out of A total possible actions as defined by the environment/problem), while the observation prediction layer $\mathbf{z}^3_o \in \mathcal{R}^{D\times 1}$ contains $D = 4$ neurons, making it the same dimensionality/shape as the observation space.

Learning While there are many possible ways to adjust the values inside of Θ, we opted to design a cost function and calculate the gradients of this objective with respect to the synaptic weight matrices of our model for the sake of simulation speed. The cost function that we designed to train our full agent was multi-objective in nature and is defined in the following manner:

$$\mathcal{L}(\mathbf{o}_{t+1}, \mathbf{t}; \Theta) = \mathcal{L}_a(\mathbf{o}_{t+1}; \Theta) + \mathcal{L}_o(\mathbf{t}_{t+1}; \Theta) \tag{5}$$

$$\mathcal{L}_a(\mathbf{t}; \Theta) = \frac{1}{2\sigma_a^2}||\mathbf{t} - \mathbf{z}^3_a||^2_2 \tag{6}$$

$$\mathcal{L}_o(\mathbf{o}_{t+1}; \Theta) = \frac{1}{2\sigma_o^2}||\mathbf{o}_{t+1} - \mathbf{z}^3_o||^2_2 \tag{7}$$

Where the target value for the action output head is calculated as $t_j = r_j + \gamma \max_a f^a_\Theta(\mathbf{o}_t)$ while the target action vector is computed as $\mathbf{t}_j = t_j\mathbf{a}_j + (1 - \mathbf{a}_j) \odot f^a_\Theta(\mathbf{o}_t)$. In the above set of equations, we see that the MLP model's weights are adjusted so as to minimize the linear combination of two terms, the cost associated with the difference between a target vector $\mathbf{t}$, which contains the bootstrap-estimated of the EFE values, and the agent's original estimate $\mathbf{z}^3_a$ as well as the cost associated with how far off the agent's prediction/expectation $\mathbf{z}^3_o$ of its environment is from the actual observation $\mathbf{o}_{t+1}$. In this study, the standard deviation coefficients associated with both output layers are set to one, i.e., $\sigma_a = \sigma_o = 1$ (highlighting that we assume unit variance for our model's free energy estimates and its environmental state predictions—note that a dynamic variance could be modeled by adding an additional output head responsible for computing the aleatoric uncertainty associated with $\mathbf{o}_{t+1}$).

Updating the parameters Θ of the neural system then consists of computing the gradient $\frac{\partial \mathcal{L}(\mathbf{o}_{t+1}, \mathbf{t}; \Theta)}{\partial \Theta}$ using reverse-mode differentiation and adjusting their values using a method such as stochastic gradient descent or variants, e.g., Adam (Kingma and Ba, 2014), RMSprop (Tieleman et al., 2012). Specifically, at each time step of any simulated episode, our agent first stores the current transition of the form $(\mathbf{o}_t, \mathbf{a}_t, r_t, \mathbf{o}_{t+1})$ into an episodic memory replay buffer (Mnih et al., 2015) and then immediately calculates $\frac{\partial \mathcal{L}(\mathbf{o}_{t+1}, \mathbf{t}; \Theta)}{\partial \Theta}$ from a batch of observation/transition data (uniformly) sampled from the replay buffer, which stores up to 10^5 transitions. We will demonstrate the benefit of this design empirically in the results section.

3. Results

3.1. Hypotheses for interception strategies

Given the fact that the target changes its speed during an episode in our interception task, the agent / human subject could gain advantage by anticipating the target speed change prior to the change of target speed. To select an optimal action early within the trail, the agent needs to take into consideration the initial target speed in the current episode and make adjustments based on the experience acquired from previous episodes. So, the question becomes: how does the agent adapt its behavior on the basis of current episode's observation of target speed/distance from the interception point and the learned statistics across episodes?

3.2. Experimental setup

We implemented the interception task as an environment in Python based on the OpenAI gym (Brockman et al., 2016) library. This integration provides the full functionality and usability of the gym environment, which means that the environment can work / be used with any RL algorithm and is made accessible to the machine learning community as well. Our AIF agents and baseline algorithm DQN are implemented with the Tensorflow2 (Abadi et al., 2015) library. Experimental data and code will be made publicly available upon acceptance.

3.3. Task performance

We compare AIF agents with and without the *epistemic* component and a baseline algorithm, i.e., a deep-Q network (DQN) (Mnih et al., 2015). We define a trial as a computational experiment where the agent performs the interception task sequentially for a number of episodes. We run a number of trials and then calculate the mean and standard deviation across trials in order to obtain a statistically valid results. The simulations in our study set the update frequency of the task environment to be $60Hz$ in order to match the exact frequency of the original human study by Diaz et al. (2009). During each episode, the joint model receives an observation each time step at $60Hz$ and estimates the EFE value for each possible action. Finally, an action is selected based on the estimated EFE values and executed in the environment. This process corresponds to Section 2.4.2. Experiments are conducted for 20 trials where each trial contains 3000 episodes. The task performance of agents is shown as curves plotting window-averaged rewards (with a window size of 100 episodes) in Figure 4, where the solid line depicts the mean value across trials and the shaded area represents standard deviation. We conducted a set of experiments where the discount factor γ of the models and the pedal lag coefficient K were varied (note that, in AIF and RL research, γ is typically fixed to a value between 0.9 and 1 to enable the model to account for long term returns). In order to compare the performance of our agents to that of human subjects, we apply the original pedal lag coefficient in one set of our experiments (specifically shown in Figure 4C).

Observe that our AIF agents are able to reach around a 90% success rate stably with very low variance. This beats human performance with 47% (std = 11.31) on average and 54.9% in the final block of experiments reported in Diaz et al. (2009). The baseline DQN agent, which learns from the problem's sparse reward signal at the end of each episode, yields an average success rate of 22% at test time. Similarly, the AIF agent with both *instrumental* and *epistemic* components achieves a 90% mean success rate.

Note that the DQN agent is outperformed by the AIF agents trained with our customized prior preference function by a large margin. This reveals that the flexibility of injecting prior knowledge is crucial for solving complex tasks more efficiently and validates our motivation of applying AIF to cognitive tasks. In our preliminary experiments, we tested an AIF agent which consists of an EFE network and a transition network separately. This AIF agent is outperformed by the AIF agent with joint model in terms of windowed mean rewards and stability. Furthermore, the AIF agent with joint model has lower model complexity. Specifically, AIF agent with joint model has only 66.8% of the parameter counts of that of AIF agent with separate models. This supports our intuition that combining the EFE model with the transition model yields an overall better model agent.

Interestingly, the AIF agent with only *instrumental* component was able to nearly reach the same level of performance as the full AIF agent. However, success rate of this agent exhibited a larger variance than the full AIF agent. Based on comparison between agents with and without *epistemic* component, we argue that *epistemic* component serves, at least in the context of the interception task we investigate, as a regularizer for the AIF models, providing improved robustness. Since we apply experience replay and bootstrapping to train the AIF models, it is possible that a local minimum is reached in the optimization process because the replay buffer is filled up with samples which come from the same subspace as the state space. Therefore, with the help of *epistemic* component, the agent is encouraged to explore the environment more often and adjusts its prediction of future observations such that it has a higher chance of escaping poorer local optima. Our proposed AIF agent reaches a plateau in performance after about $1,000$ episodes and stabilizes more after $1,500$ episodes. Note that, in contrast, human subjects were able to perform the task at an average success rate after 9 episodes of initial practice (Diaz et al., 2009).

3.4. Anticipatory behavior of AIF agents

Do the AIF agents exhibit a similar capacity for anticipatory behavior as humans do? To answer this question and to compare the strategy used by our AIF agents to that of human subjects, we record the Time-To-Contact (TTC) from trained AIF agents at the onset of the target's speed change in each episode. We then calculate,

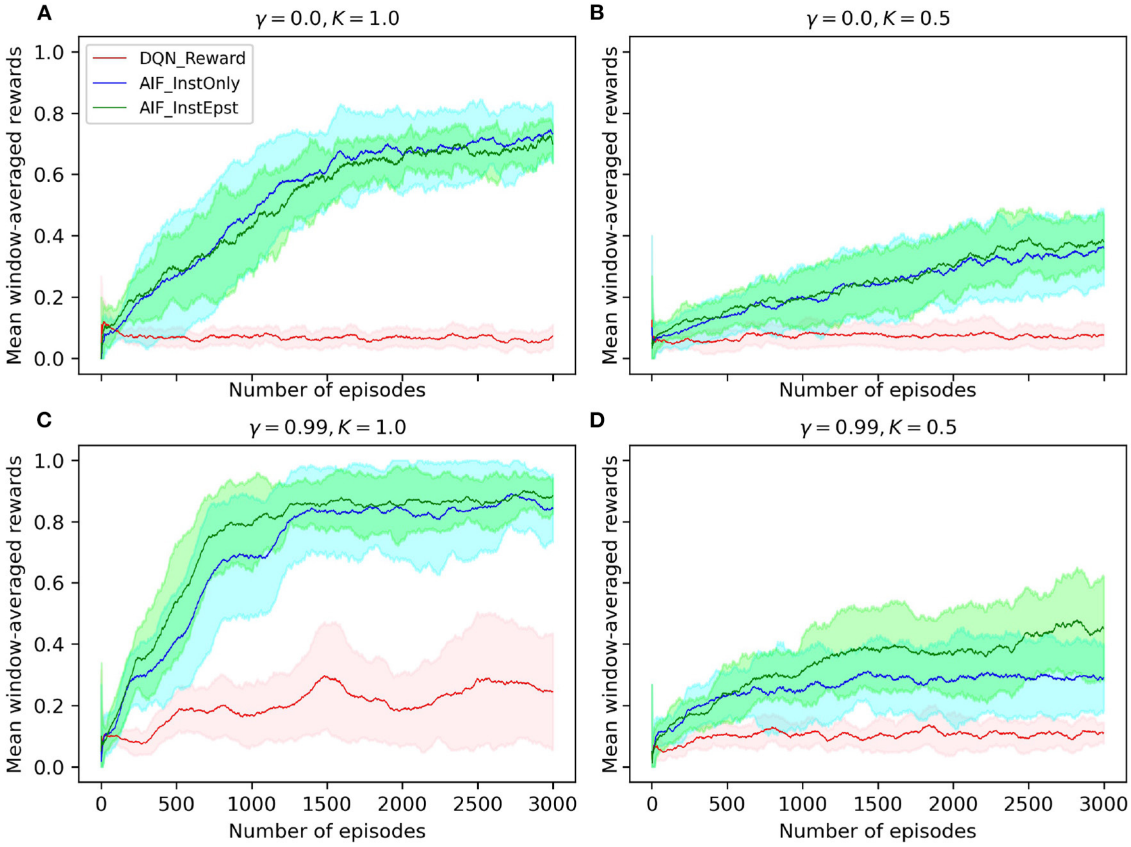

FIGURE 4
(A–D) Window-averaged reward measurements of agent performance on the interception task. *DQN_Reward* represents a DQN agent that utilized the sparse reward signal and ϵ − *greedy* exploration; *AIF_InstOnly* represents our AIF agent with only *instrumental* component which is defined by the *prior mapping function*; *AIF_InstEpst* represents an AIF agent that consists of both *instrumental* and *epistemic* components. Discount factor is denoted by γ, pedal lag coefficient is denoted by K.

at the same time: 1) the target's TTC using first-order information, and 2) target's TTC with the assumption that the target would change its speed at the most likely time and reach an averaged final speed. Finally, we compose these three types of TTC data grouped by target initial speed into a single boxplot in Figure 5. Following the assumptions made in Diaz et al. (2009), we expect that the agent would adjust its speed in a way such that its first-order TTC will equal the target's first-order TTC before it learns enough from experience to realize that the target almost always accelerates. The target's actual TTC with the interception point would be less than the first-order TTC if the target accelerates midway through. If the agent is able to anticipate the target's acceleration later in the episode, it should accelerate even before the target does in order to match the target's actual TTC with the interception point.

In our experimental analysis, we found that the discount factor γ plays a big role in forming different behavior patterns within AIF agents. All variants of AIF agents were trained with the *instrumental* value computed using our first-order *prior mapping function*. Intuitively, the agent's behavior should conform to a reactive agent who uses only the first-order information and acts to match its own TTC to the target's first-order TTC, just like what has been observed in Figure 5A (please see that the green box is nearly identical to the blue box under all target initial conditions). The AIF agent depicted in Figure 5A is set to use a discount factor of 0, which means that the agent only seeks to maximize its immediate reward without considering the long-term impact of the action(s) that it selects. Such an agent converges to a reactive behavior. However, when we increase the discount factor to 0.99 (which is a common practice in RL literature), the AIF agent starts to behave more interestingly. In Figure 5C, the agent's TTC (green box) lies in between target's first-order TTC (blue box) and target's actual mean TTC (orange box), which suggests that the AIF agent tends to move faster than a pure-reactive, first-order agent would in the early phase of interception. In other words, the agent tends to anticipate the likely target speed change in the future and adjusts its action selection policy. This behavioral pattern can be explained as exploiting the benefits provided by estimating long-term accumulated *instrumental* reward signal (when the discount factor value is increased). Given a higher discount factor, in this case $\gamma = 0.99$, the AIF agent estimates the summation of *instrumental* values from its current (time) step in the task until the end of the interception using discounting. This leads to an agent who seeks to maximize long-term benefits in terms of reaching the goal when selecting actions.

3.5. Effect of vehicle dynamics on agent behavior

To test how anticipatory behavior is affected when simple reactive behavior is no longer sufficient, we increased the inertia on the

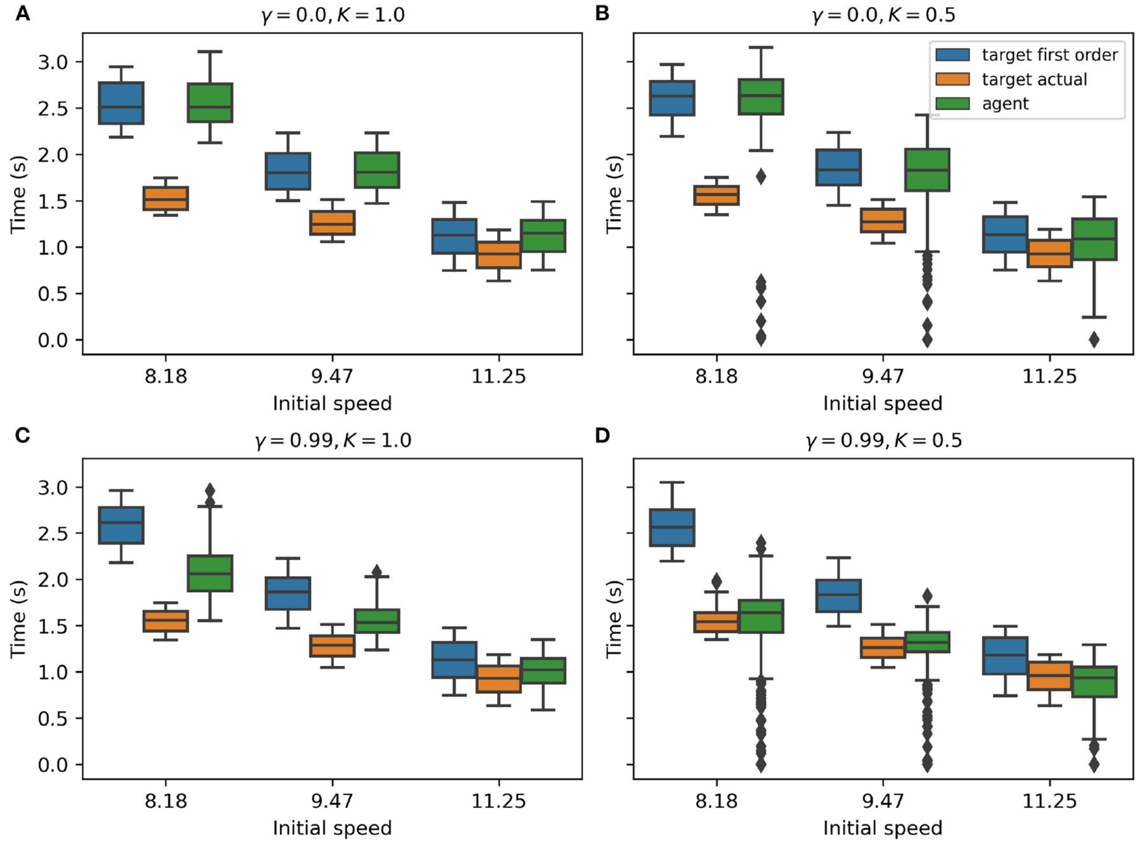

FIGURE 5
(A–D) TTC values taken at the onset of target's speed change. In each subplot, the target's first-order TTC, the target's actual mean TTC, and the agent's TTC are shown in different colors, with data grouped by target initial speed. The discount factor is denoted by γ while the pedal lag coefficient is denoted by *K*.

agent's vehicle by changing the pedal lag coefficient K. Given the same discount factor $\gamma = 0.99$, we compare two different pedal lag coefficients $K = 1.0$ in Figure 5C and $K = 0.5$ in Figure 5D, where lower K indicates less responsive vehicle dynamics. With the same discount factor, the AIF agent performing the task under a lower pedal lag coefficient in Figure 5D has a lower success rate in intercepting the target. This is due to the fact that the agent's ability to manipulate its own speed is limited, therefore there is less room left for error. However, the AIF agent in this condition yields TTC values that are closer to the target's actual mean TTC. Note that, when the target initial speed is 11.25 m/s (Figure 5D), the median of agent's TTC value is actually smaller than target's actual mean TTC. This supports our hypothesis that purely reactive behavior is not sufficient for successful interception and anticipatory behavior is emergent when the vehicle becomes less responsive.

4. Discussion

Variations of an AIF agent were trained to manipulate the speed of movement so as to intercept a target moving across the ground plane, and eventually across the agent's linear path of travel. On each episode, the target would change in speed on most episodes to a value that was selected from a Gaussian distribution of final speeds. The results demonstrate that the AIF framework is able to model both on-line visual and anticipatory control strategies in an interception task, as was previously demonstrated by humans performing the same task (Diaz et al., 2009). The agent's anticipatory behavior aimed to maximize the cumulative expected free energy in the duration that follows action selection. Variation of the agent's discount factor modified the length of this duration. At lower discount factors, the agent behaved in a reactive manner throughout the approach, consistent with the constant bearing angle strategy of interception. At higher values, actions that were selected before the predictable change in speed took into account the most likely change in target speed that would occur later in the episode. Anticipatory behavior was also influenced by the agent's capabilities for action.This anticipatory behavior was most apparent when the pedal lag coefficient was set to lower values, which had the effect of changing the agent's movement dynamics so that purely reactive control was insufficient for interception behavior.

Despite the agent's demonstration of qualitatively human-like prediction, careful comparison of the agent's behavior to the human performance and learning rates demonstrated in Diaz et al. (2009) reveals notable differences. Analysis of participant behavior in the fourth and final block of Experiment 1 in Diaz et al. (2009) reveals

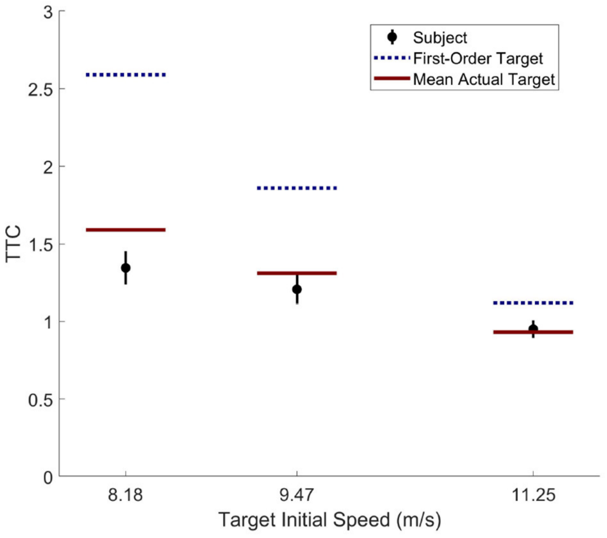

FIGURE 6
Human subject data from Exp 1. of Diaz et al. (2009). TTCs were taken at the onset of target's speed change. Dotted line represents the mean of target's first-order TTC, solid line represents the mean of target's actual TTC, black disk represents the mean of subject's TTC with a bar indicating 95% confidence interval of the mean.

that subject TTC at the onset of the target's change in speed was well matched to the most likely time and magnitude of the target's likely change in speed (i.e., the mean actual target TTC in Figure 6). In contrast, the AIF agent with an equivalent pedal lag ($K = 1.0$; i.e., the *matched* agent) demonstrated only partial matching of its TTC to the likely change in target speed (the target's mean actual TTC in Figure 5C). Although one might attribute this to under-training of the agent, it is notable that the agent achieved a hit rate exceeding 80% by the end of training, while human participants in the original study consistently improved in performance until reaching 55% hit rate at the end of the experiment.

To better understand the potential causes of these differences between agent and human performance, it is helpful to consider how the agent's mechanism for anticipation differs from that of humans. The agent chooses actions on the basis of a weighted combination of reward-based reinforcement (instrumental reward) and short model-based prediction (epistemic reward), both of which are computed within the two-headed joint model. EFE values are computed in the EFE head, which is responsible for selecting the action (i.e., pedal position) that it estimates would produce the lowest expected free energy later in the agent's approach. The estimate of EFE associated with each pedal position does not involve an explicit process of model-based prediction, but is learned retrospectively, through the use of an experience replay buffer. Following action selection, visual feedback provides an indication of the cumulative EFE over the duration of the replay buffer. The values of EFE within this buffer are weighted by their temporal distance from the selected action in accordance with the parameter of discount factor. This is similar to both reward-based learning and is often compared to the dopaminergic reward system in humans (Holroyd and Coles, 2002; Haruno, 2004; Lee et al., 2012; Momennejad et al., 2017). The epistemic component of the EFE reward signal is thought to drive exploration toward uncertain world states, and it relies on predictions made in the transition head. This component of the model relies on the hidden states provided by the shared neural layers in the joint model and predicts an observation at next time step $\hat{\mathbf{o}}_{t+1}$. The estimated observation at next time step is then compared to the ground truth observation $\mathbf{o}_{t+1}$ and the difference between them generates the epistemic signal $R_{t,e}$. The role of the transition head is in many ways consistent with a "strong model-based" form of prediction (Zhao and Warren, 2015), whereby predictive behaviors are planned on the basis of an internal model of world states and dynamics that facilitate continuous extrapolation. In summary, whereas the EFE head is consistent with reward based learning, the transition head is consistent with relatively short-term model based prediction.

How does this account of anticipation demonstrated by our agent compare with what we know about anticipation in humans? As discussed in the introduction, empirical data on the quality of model-based prediction suggests that it degrades sufficiently quickly that it cannot explain behaviors of the sort demonstrated here, by our agent, or by the humans in Diaz et al. (2009). In contrast, a common theory in motor control and learning relies upon a comparison of a very short-term prediction (e.g., milliseconds) of self-generated action with immediate sensory feedback (Hoist et al., 1950; Wade, 1994; Wolpert et al., 1995; Blakemore et al., 1998). However, this similarity is weakened by the observation that, in the context of motor-learning, short-term prediction is thought to rely upon access to an efferent copy of the motor signal used to generate the action. For this reason, it is problematic that the AIF agent is predicting both its own future state (x_s, v_s) and the future state of the target (x_t, v_t), for which there is no efferent copy or analogous information concerning movement dynamics. Although research on eye movements has revealed evidence for the short-term prediction of future object position and trajectory (Ferrera and Barborica, 2010; Diaz et al., 2013a,b), it remains unclear whether these behaviors are the result of predictive models of object dynamics or representation-minimal heuristics.

Another possible contribution to the observed differences between agent and human performance is the perceptual input. When considering potential causes for the difference between agent and human anticipatory behavior, it is notable that the agent relies upon an observation vector defined by agent's and target's position and velocity measured in meters, and meters per second, respectively. However, in the natural context, these spatial variables must be recovered or estimated on the basis of perceptual sources of information, such as the rate of global optic flow due to translation over the ground plane, the exocentric direction of the target, the instantaneous angular size of the target, or the looming rate of the target during the agent's approach. It is possible that by depriving the agent of these optical variables, we are also depriving the agent of opportunities to exploit task-relevant relationships between the agent and environment, such as the bearing angle. It is also notable that some perceptual variables may provide redundant information about a particular spatial variable (e.g., both change in bearing angle and rate of change in angular size may be informative about an objects approach speed). However, redundant variables will differ in reliability by virtue of sensory thresholds and resolutions. For these reasons, a more complete and comprehensive model of human visually guided action and anticipation would take as input potential sources of information and learn to weight them according to context-dependent reliability and variability.

Another potential contributor to differences between human and agent performance is the notable lack of visuo-motor delays within the agent's architecture. In contrast, human visuo-motor delay has been estimated to be on the order of 100–200 ms between the arrival of new visual information and the modification or execution of an action (Nijhawan, 2008; Le Runigo et al., 2010). Because uncompensated delays would have devastating consequences on human visual and motor control, they are often cited as evidence that humans must have some form of predictive mechanism that acts in compensation (Wolpert et al., 1995). Future attempts to make this model's anticipatory behavior more human-like in nature may do so by imposing similar length delay between the agent's choice of motor plan on the basis of the observed world-state and the time that this motor plan is executed (Walsh et al., 2009). Finally, note that our proposed architecture is "flat" in the temporal sense. It other words, EFE values are calculated and actions are planned in a single linear time scale. In contrast, a deep/hierarchical temporal model would imply that policies are inferred, learned, and ultimately operate at different time scales (Friston et al., 2018). We believe that our approach is sufficient for the given task of this study. However, if one intended to extend the problem to more sophisticated settings where higher level cognitive functions are separated from lower-level motor control, a deep temporal model could be a more suitable/useful approach.

Due to limited computation resources that we have access to and the high computational cost of the full Bayesian inference framework (which, in the context of neural networks, requires formulating each neural network as a Bayesian neural network where training, typically to obtain good-quality performance, requires Markov chain Monte Carlo), we simplify the Bayesian inference by assuming a uniform prior (or uninformative prior) on the parameters of our model, similar to Tschantz et al. (2020a). Maximum likelihood estimation (MLE), in our setup, is generally equivalent to maximum a posteriori (MAP) estimation while assuming the priors to be uniform distributions. More general forms of Bayesian inference with different prior assumptions could be examined in future work. Also, note that the Laplace approximation applied in this work leads to the expected free energy reducing to a KL-divergence (i.e., KL control).

5. Conclusion

We present a novel scaled-up version of active inference framework (AIF) model for studying online visually guided locomotion using an interception task where a moving target changes its speeds in a semi-predictable manner. In order to drive the agent toward the goal more effectively, we devised a problem-specific *prior mapping function*, improving the agent's computational efficiency and interpretability. Notably, we found that our proposed AIF agent exhibits better task performance when compared to a commonly used RL agent, i.e., the deep-Q network (DQN). The full AIF agent, containing both *instrumental* and *epistemic* components, exhibited slightly better task performance and lower variance compared to the AIF agent with only an *instrumental* component. Furthermore, we demonstrated behavioral differences among our full AIF agents given different discount factor γ values as well as levels of the agent's action-to-speed responsiveness. Finally, we analyzed the anticipatory behavior demonstrated by our agent and examined the differences between the agent's behavior and human behavior. While our results are promising, future work should address the following limitations—first, inputs to our agent are defined in a simplified vector space whereas sensory inputs to the humans that actually perform the interception task are visual in nature (i.e., the model should work directly with unstructured sensory data such as pixel values). We remark that a vision-based approach could facilitate the extraction of additional information and features that are useful for solving the interception task more reliably. Second, our simulations do not account for visuo-motor delays inherent to the human visual and motor systems, and that might be modeled using techniques like delayed Markov decision process formulations (Walsh et al., 2009; Firoiu et al., 2018).

Data availability statement

The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Ethics statement

The Diaz et al. (2009) study was reviewed and approved by Rensselaer Polytechnic Institute IRB. The patients/participants provided their written informed consent to participate in that study. Ethical review and approval was not required for the the present study in accordance with the local legislation and institutional requirements. Written informed consent to participate in the current study was not required in accordance with the local legislation and institutional requirements.

Author contributions

AO aided ZY in preliminary simulation/testing and they both devised the neural AIF algorithm. ZY implemented the experimental simulations as well as collected and analyzed the results. All authors contributed to the experimental design and the project's development, data interpretation, drafting of the manuscript, and approval of the final version of the manuscript for submission.

Funding

This material was based upon work supported by the National Science Foundation (NSF) under Award No. DGE-2125362 and CPS-2225354.

Acknowledgments

The authors would like to thank Tim Johnson for contributing to the initial software development of the interception task simulation environment(s) used in this project.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Available online at: https://www.tensorflow.org

Blakemore, S.-J., Wolpert, D. M., and Frith, C. D. (1998). Central cancellation of self-produced tickle sensation. *Nat. Neurosci.* 1, 635–640. doi: 10.1038/2870

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., et al. (2016). OpenAi gym. *arXiv preprint* arXiv:1606.01540. doi: 10.48550/arXiv.1606.01540

Buckley, C. L., Kim, C. S., McGregor, S., and Seth, A. K. (2017). The free energy principle for action and perception: a mathematical review. *J. Math. Psychol.* 81, 55–79. doi: 10.1016/j.jmp.2017.09.004

Çatal, O., Wauthier, S., De Boom, C., Verbelen, T., and Dhoedt, B. (2020). Learning generative state space models for active inference. *Front. Comput. Neurosci.* 14, 574372. doi: 10.3389/fncom.2020.574372

Chapman, S. (1968). Catching a baseball. *Am. J. Phys.* 36, 868–870. doi: 10.1119/1.1974297

Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., and Friston, K. (2020). Active inference on discrete state-spaces: a synthesis. *J. Math. Psychol.* 99, 102447. doi: 10.1016/j.jmp.2020.102447

Diaz, G., Cooper, J., and Hayhoe, M. (2013a). Memory and prediction in natural gaze control. *Philos. Trans. R. Soc. B Biol. Sci.* 368, 20130064. doi: 10.1098/rstb.2013.0064

Diaz, G., Cooper, J., Rothkopf, C., and Hayhoe, M. (2013b). Saccades to future ball location reveal memory-based prediction in a virtual-reality interception task. *J. Vis.* 13, 20–20. doi: 10.1167/13.1.20

Diaz, G. J., Parade, M. S., Barton, S. L., and Fajen, B. R. (2018). The pickup of visual information about size and location during approach to an obstacle. *PLoS ONE* 13, e0192044. doi: 10.1371/journal.pone.0192044

Diaz, G. J., Phillips, F., and Fajen, B. R. (2009). Intercepting moving targets: a little foresight helps a lot. *Exp. Brain Res.* 195, 345–360. doi: 10.1007/s00221-009-1794-5

Fajen, B., Diaz, G., and Cramer, C. (2008). Reconsidering the role of action in perceiving the catchability of fly balls. *J. Vis.* 8, 621–621. doi: 10.1167/8.6.621

Fajen, B. R., and Devaney, M. C. (2006). Learning to control collisions: the role of perceptual attunement and action boundaries. *J. Exp. Psychol. Hum. Percept. Perform.* 32, 300–313. doi: 10.1037/0096-1523.32.2.300

Fajen, B. R., and Warren, W. H. (2007). Behavioral dynamics of intercepting a moving target. *Exp. Brain Res.* 180, 303–319. doi: 10.1007/s00221-007-0859-6

Ferrera, V. P., and Barborica, A. (2010). Internally generated error signals in monkey frontal eye field during an inferred motion task. *J. Neurosci.* 30, 11612–11623. doi: 10.1523/JNEUROSCI.2977-10.2010

Firoiu, V., Ju, T., and Tenenbaum, J. (2018). At human speed: deep reinforcement learning with action delay. *arXiv preprint* arXiv:1810.07286. doi: 10.48550/arXiv.1810.07286

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci.* 13, 293–301. doi: 10.1016/j.tics.2009.04.005

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. (2019). A free energy principle for a particular physics. *arXiv preprint* arXiv:1906.10184. doi: 10.48550/arXiv.1906.10184

Friston, K., Da Costa, L., Hafner, D., Hesp, C., and Parr, T. (2021). Sophisticated inference. *Neural Comput.* 33, 713–763. doi: 10.1162/neco_a_01351

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017). Active inference: a process theory. *Neural Comput.* 29, 1–49. doi: 10.1162/NECO_a_00912

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., and Pezzulo, G. (2015). Active inference and epistemic value. *Cogn. Neurosci.* 6, 187–214. doi: 10.1080/17588928.2015.1020053

Friston, K., Samothrakis, S., and Montague, R. (2012). Active inference and agency: optimal control without cost functions. *Biol. Cybern.* 106, 523–541. doi: 10.1007/s00422-012-0512-8

Friston, K. J., Daunizeau, J., and Kiebel, S. J. (2009). Reinforcement learning or active inference? *PLoS ONE* 4, e6421. doi: 10.1371/journal.pone.0006421

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., and Ondobaka, S. (2017). Active inference, curiosity and insight. *Neural Comput.* 29, 2633–2683. doi: 10.1162/neco_a_00999

Friston, K. J., Rosch, R., Parr, T., Price, C., and Bowman, H. (2018). Deep temporal models and active inference. *Neurosci. Biobehav. Rev.* 90, 486–501. doi: 10.1016/j.neubiorev.2018.04.004

Haruno, M. (2004). A neural correlate of reward-based behavioral learning in caudate nucleus: a functional magnetic resonance imaging study of a stochastic decision task. *J. Neurosci.* 24, 1660–1665. doi: 10.1523/JNEUROSCI.3417-03.2004

Hoist, E., v., Mittelstaedt, H., and Martin, R. (1950). Das reafferenzprìnzip. wechselwirkung zwischen zentralnervensystem und peripherie. *Die Naturwissenschaften* 37, 464. doi: 10.1007/BF00622503

Holroyd, C. B., and Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychol. Rev.* 109, 679–709. doi: 10.1037/0033-295X.109.4.679

Kaplan, R., and Friston, K. J. (2018). Planning and navigation as active inference. *Biol. Cybern.* 112, 323–343. doi: 10.1007/s00422-018-0753-2

Kingma, D., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint* arXiv:1412.6980. doi: 10.48550/arXiv.1412.6980

Le Runigo, C., Benguigui, N., and Bardy, B. G. (2010). Visuo-motor delay, information-movement coupling, and expertise in ball sports. *J. Sports Sci.* 28, 327–337. doi: 10.1080/02640410903502782

Lee, D., Seo, H., and Jung, M. W. (2012). Neural basis of reinforcement learning and decision making. *Annu. Rev. Neurosci.* 35, 287–308. doi: 10.1146/annurev-neuro-062111-150512

Lee, D. N. (1976). A theory of visual control of braking based on information about time-to-collision. *Perception* 5, 437–459. doi: 10.1068/p050437

Matthis, J. S., and Fajen, B. R. (2013). Humans exploit the biomechanics of bipedal gait during visually guided walking over complex terrain. *Proc. R. Soc. B Biol. Sci.* 280, 20130700. doi: 10.1098/rspb.2013.0700

Millidge, B. (2020). Deep active inference as variational policy gradients. *J. Math. Psychol.* 96, 102348. doi: 10.1016/j.jmp.2020.102348

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. *Nature* 518, 529–533. doi: 10.1038/nature14236

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., and Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nat. Hum. Behav.* 1, 680–692. doi: 10.1038/s41562-017-0180-8

Nijhawan, R. (2008). Visual prediction: Psychophysics and neurophysiology of compensation for time delays. *Behav. Brain Sci.* 31, 179–198. doi: 10.1017/S0140525X08003804

Parr, T., and Friston, K. J. (2017). Uncertainty, epistemics and active inference. *J. R. Soc. Interface* 14, 20170376. doi: 10.1098/rsif.2017.0376

Parr, T., and Friston, K. J. (2018). The anatomy of inference: generative models and brain structure. *Front. Comput. Neurosci.* 12, 90. doi: 10.3389/fncom.2018.00090

Parr, T., and Friston, K. J. (2019). Generalised free energy and active inference. *Biol. Cybern.* 113, 495–513. doi: 10.1007/s00422-019-00805-w

Sajid, N., Ball, P. J., Parr, T., and Friston, K. J. (2021). Active inference: demystified and compared. *Neural Comput.* 33, 674–712. doi: 10.1162/neco_a_01357

Schwartenbeck, P., Passecker, J., Hauser, T. U., FitzGerald, T. H., Kronbichler, M., and Friston, K. J. (2019). Computational mechanisms of curiosity and goal-directed exploration. *Elife* 8, e41703. doi: 10.7554/eLife.41703

Shin, J. Y., Kim, C., and Hwang, H. J. (2022). Prior preference learning from experts: designing a reward with active inference. *Neurocomputing* 492, 508–515. doi: 10.1016/j.neucom.2021.12.042

Tieleman, T., Hinton, G., et al. (2012). Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. *Coursera Neural Netw. Mach. Learn.* 4, 26–31.

Tschantz, A., Baltieri, M., Seth, A. K., and Buckley, C. L. (2020a). "Scaling active inference," in *2020 International Joint Conference on Neural Networks (IJCNN)* (Glasgow, UK: IEEE), 1–8.

Tschantz, A., Seth, A. K., and Buckley, C. L. (2020b). Learning action-oriented models through active inference. *PLoS Comput. Biol.* 16, e1007805. doi: 10.1371/journal.pcbi.1007805

Ueltzhöffer, K. (2018). Deep active inference. *Biol. Cybern.* 112, 547–573. doi: 10.1007/s00422-018-0785-7

Wade, N. J. (1994). Hermann von helmholtz (1821–1894). *Perception* 23, 981–999. doi: 10.1068/p230981

Walsh, T. J., Nouri, A., Li, L., and Littman, M. L. (2009). Learning and planning in environments with delayed feedback. *Auton. Agent Multi Agent. Syst.* 18, 83–105. doi: 10.1007/s10458-008-9056-7

Warren, W., Fajen, B., and Belcher, D. (2010). Behavioral dynamics of steering, obstacle avoidance, and route selection. *J. Vis.* 1, 184–184. doi: 10.1167/1.3.184

Wolpert, D. M., Ghahramani, Z., and Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science* 269, 1880–1882. doi: 10.1126/science.7569931

Yilmaz, E. H., and Warren, W. H. (1995). Visual control of braking: a test of the hypothesis. *J. Exp. Psychol. Hum. Percept. Perform.* 21, 996. doi: 10.1037/0096-1523.21.5.996

Zhao, H., and Warren, W. H. (2015). On-line and model-based approaches to the visual control of action. *Vis. Res.* 110, 190–202. doi: 10.1016/j.visres.2014.10.008