Toward Human-AI Alignment in Large-Scale Multi-Player Games

Sugandha Sharma[1,2],
Guy Davidson[1,3], **Khimya Khetarpal**[1,4], **Anssi Kanervisto**[1], **Udit Arora** [1],
Katja Hofmann[1], **Ida Momennejad**[1],

[1]Microsoft Research, [2]Massachusetts Institute of Technology,
[3]New York University, [4]Mila, University of Montreal
Correspondence: susharma@mit.edu

Published in Wordplay @ ACL 2024, Association for Computational Linguistics

Abstract

Achieving human-AI alignment in complex multi-agent games is crucial for creating trustworthy AI agents that enhance gameplay. We propose a method to evaluate this alignment using an interpretable task-sets framework, focusing on high-level behavioral tasks instead of low-level policies. Our approach has three components. First, we analyze extensive human gameplay data from Xbox's Bleeding Edge (100K+ games), uncovering behavioral patterns in a complex task space. This task space serves as a basis set for a behavior manifold capturing interpretable axes: fight-flight, explore-exploit, and solo-multi-agent. Second, we train an AI agent to play Bleeding Edge using a Generative Pretrained Causal Transformer and measure its behavior. Third, we project human and AI gameplay to the proposed behavior manifold to compare and contrast. This allows us to interpret differences in policy as higher-level behavioral concepts, e.g., we find that while human players exhibit variability in fight-flight and explore-exploit behavior, AI players tend towards uniformity. Furthermore, AI agents predominantly engage in solo play, while humans often engage in cooperative and competitive multi-agent patterns. These stark differences underscore the need for interpretable evaluation, design, and integration of AI in human-aligned applications. Our study advances the alignment discussion in AI and especially generative AI research, offering a measurable framework for interpretable human-agent alignment in multiplayer gaming.

1 Introduction

Human-AI alignment is pivotal in generative AI research for several compelling reasons. First, as generative AI is increasingly integrated into various applications (Park et al., 2023; Brynjolfsson et al., 2023), ensuring alignment with human values and intentions becomes crucial to mitigate risks and enhance user trust (Sucholutsky et al., 2023; Gabriel, 2020). Second, aligning AI systems with human behavior fosters more effective collaboration between humans and machines (Chakraborti and Kambhampati, 2018), unlocking the potential for synergistic outcomes (Wynn et al., 2023; Bobu et al., 2023). Third, in ethical considerations surrounding the deployment of generative models, alignment serves as a safeguard against unintended consequences and biases, promoting responsible AI development and deployment (Kenthapadi et al., 2023; Weidinger et al., 2021). Thus, human-AI alignment in generative AI is essential for creating trustworthy, beneficial, and ethically sound AI applications that align with societal values and expectations.

The evaluation of human-agent alignment in rich observation and multi-agent environments poses a significant challenge (Leike et al., 2018; Ouyang et al., 2022; Wang et al., 2022, 2023b; Burns et al., 2023). In complex multi-agent video games, where each player faces a multitude of actions, this challenge involves finding the appropriate level of abstraction for a meaningful interpretation of human actions to evaluate artificial agents' alignment.

In this work, we propose an interpretable approach towards human-AI alignment by introducing the “Task-sets" framework, offering a means to abstract task sets from the environment. Task-sets (Sakai, 2008) offer a higher level of abstraction compared to policies (Sutton et al., 1999a; Silver et al., 2014; Lillicrap et al., 2015; Schulman et al., 2017) and options (Sutton et al., 1999b; Precup, 2000; Stolle and Precup, 2002; Bacon et al., 2017; Khetarpal et al., 2020) in reinforcement learning.

An agent's policy provides a low-level mapping from states to actions. Suppose a person wants to get a snack from their kitchen. Their policy might indicate the precise sequence of steps (actions) to take to get from their living room to their kitchen. The behavior described above might be split into two options: “go-from-livingroom-to-kitchen”, and “acquire-snack-in-kitchen”. The value of options

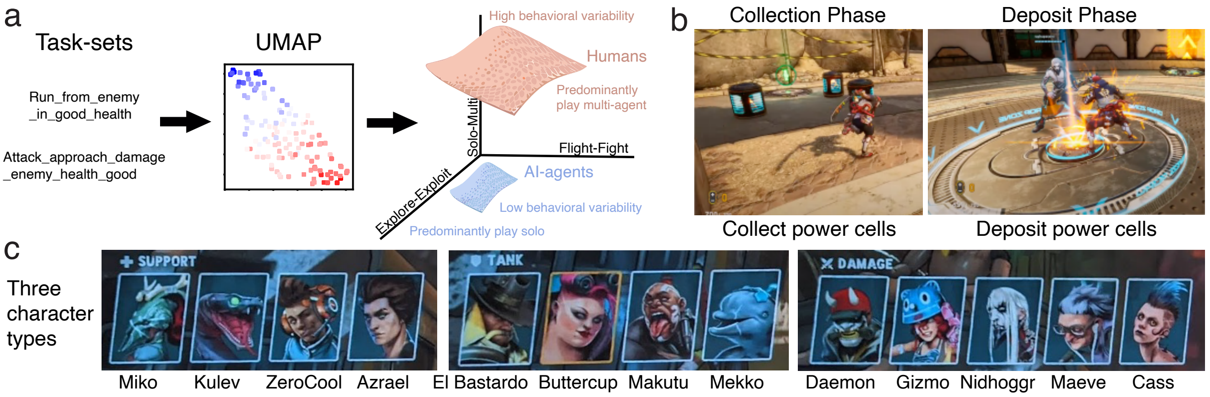


Figure 1: **Bleeding edge *Power Collection* game mode.** (a) Analysis pipeline begins with task-sets used to extract the UMAP manifold embedding, interpreted to derive 3D human and AI behavioral manifold schematic. Humans highly vary in how they express fight-flight and explore-exploit behavior; they predominantly play in a multi-agent settings. AI agents exhibit low variability in fight-flight and explore-exploit behavior tending towards uniformity; they predominantly play solo. (b) Collection phase (*left*) and Deposit phase (*right*) in the Power Collection game mode. (c) Three character types (Support, Tank and Damage) with 13 possible characters in the game.

lies in the fact that if a person is in their bedroom, they can execute a different first option (to arrive in the kitchen), and the same second option (acquiring a snack). The next day, however, at the office, their option's policy that facilitates acquiring a snack at home cannot help, as they need to navigate a different path, to a differently laid-out kitchen. Task-sets describe a general higher degree of abstraction, "walk to kitchen when hungry to find a snack," that is independent of the precise layout of the building (state space) and the steps needed to obtain the snacks (action space).

Key Definitions:

- **Task-set:** given a specific perceptual criteria (task domain) respond according to certain rules (task rules).
- **Affordance:** when the criteria for performing a task set are met. Multiple task sets can be simultaneously afforded.
- **Behavioral manifold:** dimensionality reduced space to which task set behavior and its spread are projected.
- **Alignment on the manifold:** comparing human and agent spread along the dimensions of the behavioral manifold.

We use the task-set abstraction to interpret differences between agents as higher-level behavioral concepts that transcend comparing changes in policies and options alone. This abstraction also allows comparing behavior across temporal scales (Monsell, 2003; Sakai, 2008; Collins and Frank, 2013; Momennejad and Haynes, 2013; Vaidya and Badre, 2022). Task sets also facilitate compositionality. In the example above, the "walk to kitchen when hungry to find a snack" task-set could also be used to grab a snack for a visitor (taking their preferences into account) or to modify one's own snack choice to account for healthiness. In Bleeding Edge, we use task-sets to abstract from actions and policies of players to higher-level notions, such as the dimensions of the behavior manifold (Fig.1a). This abstraction not only enables understanding of human cognitive processes, but also, fosters strong notions of transferability, both between agents and across environments, thus making it suitable for effective evaluation of alignment between AI agents and humans.

Our **key contributions** are as follows: We propose an interpretable analysis of multi-scale behavior on different tasks by projecting them to behavior manifolds (Fig.1a) and evaluating human-AI alignment in this latent space. First, we analyze human gameplay data from the Xbox game Bleeding Edge (≈ 100K games). Our analysis uncovers human behavioral patterns in a complex task-set space. We then interpret the agent's choices over which tasks to pursue at different moments in time as a behavior manifold capturing three interpretable axes: fight-flight, explore-exploit, and solo-multi-agent. Second, we train a proof of concept AI agent for gameplay using a Generative Pretrained Causal Transformer and measure its behavior using the same methods applied to the human data. Third, we project human and AI behavior to the same behavioral manifold and use the axes we defined to compare human-AI alignment. This three-fold analytical framework allows us to discern the extent of alignment between human and AI agent behaviors in a subspace defined by high-level and interpretable tasks, rather than policies.

Our research, driven by these investigative avenues, pursues two-fold primary objectives. 1) We seek a nuanced understanding of human cognition and behavior in the realm of large-scale multiplayer video games. 2) We aspire to harness this understanding to advance AI for gameplay, by constructing, evaluating and training artifical agents for targeted behavior replication through player style identification. In this work we mainly focus on the evaluation framework for measuring alignment (illustrated through a proof of concept AI agent). This framework can be used for evaluating alignment of any autonomous decision making AI agent (Yao et al., 2022; Sharma et al., 2022; Shinn et al., 2023; AutoGPT, 2023; Du et al., 2023; Wang et al., 2023a). In summary, we provide a framework to evaluate human-AI alignment that could potentially be applied towards developing AI agents with superior alignment with humans.

2 Bleeding Edge

Bleeding Edge is a dynamic and engaging large-scale multiplayer online video game developed by Ninja Theory, blending fast-paced combat mechanics with team-based strategy.

Gameplay: Bleeding Edge is designed for 4v4 multiplayer battles. This means that two teams, each consisting of four players, participate in each game. Players engage in team-based battles featuring dynamic combat dynamics, including both melee (close-quarters physical engagements) and ranged (attacks from a distance) elements.

Power Collection game mode: In this work, we restrict our analysis to the Power Collection game mode in Bleeding Edge. Central to this mode are two distinct phases: the Power Collection Phase and the Deposit Phase (Fig.1b). The objective revolves around the strategic acquisition of power cells (seeds) scattered across the game map (top-down view of a mini-map is visible to the players on the top-right corner of their screen). During the Power Collection Phase, teams are tasked with securing power cells positioned at specific locations, requiring meticulous planning and coordination. This phase introduces a dynamic interplay of risk and reward, as teams decide whether to focus on collecting cells nearer their base or venture farther into the map. The subsequent Deposit Phase involves transporting collected power cells to designated locations for scoring, further emphasizing the need for strategic decision-making and teamwork. Teams must defend their collected cells while attempting to disrupt opponents' efforts, contributing to the overall intensity and complexity of the gameplay.

Character selection and abilities: Players choose from a diverse roster of 13 characters, each with unique abilities and playstyles. Characters are classified into three categories (Fig.1c): Support, Tank, and Damage (see appendix A.5 for details).

3 AI Agent

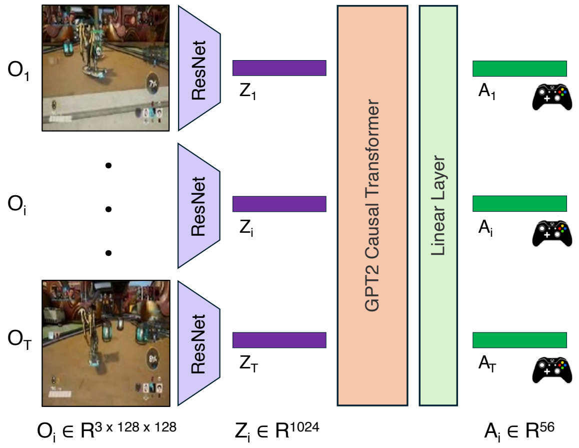


Figure 2: **AI agent architecture**. The architecture consists of a ResNet style encoder followed by a Causal Transformer. Model input is sequence of image observations (O_i), with sequence length T and the model is trained to predict actions (A_i).

We trained an AI agent for playing Bleeding Edge and measured its alignment with humans. The model, with ∼222M parameters, is a transformer based architecture. We frame human gameplay trajectories as sequences of image-action pairs, and optimize the transformer to predict the next action in the sequence given the previous images (Fig.2).

Observation Encoder: The model is trained on sequences of $T = 128$ images where each image is reshaped to $128 \times 128 \times 3$ and then divided by 255 to ensure its value lies in range [0, 1]. A custom ResNet (He et al., 2016) with 18.6M parameters

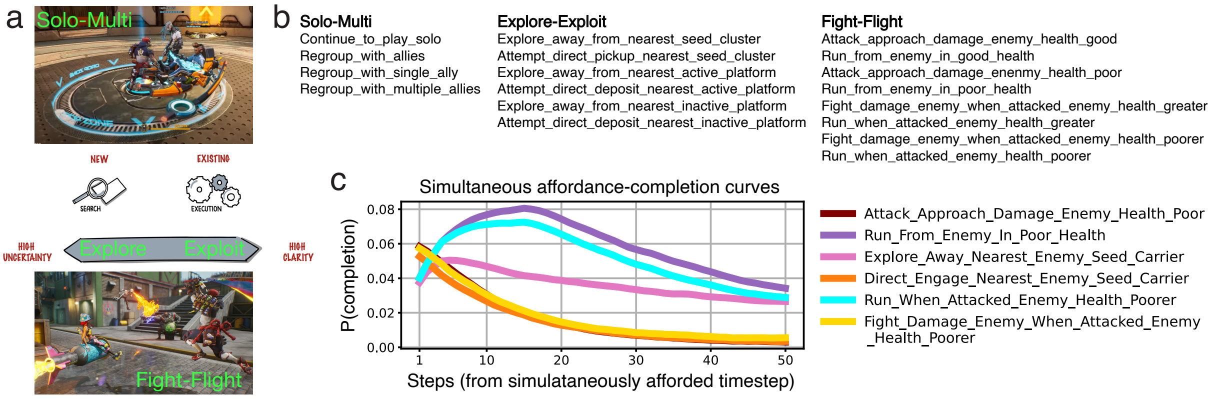


Figure 3: **Cognitive themes used for behavioral analysis.** (a) Ubiquitous cognitive themes in behaviors of various biological species used for analyzing gameplay dynamics of both human and AI agents. (b) Example task-sets defined for each of the cognitive themes. (c) Simultaneous affordance-completion curves for a subset of task-sets defined under the three cognitive themes.

is used to embed each image observation ($O \in R^{3\times128\times128}$) independently into a vector (details in A.3.2). For each input image, the output of the encoder is a 1024D embedding.

Transformer: The causal transformer (with ∼203M parameters) is applied on the image embeddings ($Z \in R^{1024}$). Specifically, a GPT2-like architecture (Radford et al., 2019) from NanoGPT (Andrej Karpathy, 2023) was used containing 16 layers/blocks. Each attention layer has 8 heads with the action embeddings output of size 1024.

Action Decoder: The final layer of the model consists of a linear layer that converts the output from the transformer (1024D) to match the dimensions of the action embedding (56D). The action space is an Xbox controller with two joysticks and 12 binary buttons. Each joystick is decomposed into x and y components leading to 4 continuous values. Each of these continuous values are discretized by binning them to 11 bins, such that the model predicts the logits over which bin is the most likely. This leads to $12 + 4 \times 11 = 56$ dimensional action output ($A \in R^{56}$).

Data and Training: We use Behavior cloning (Pomerleau, 1991) for training. For buttons we use the binary cross entropy loss with logits, and for the joysticks we use the cross entropy loss for each component. The total loss is computed as the sum of the losses for buttons and each of the joystick's components. The training data was sampled from a dataset consisting of 57,661 full human gameplay videos where each video corresponds to continuous gameplay by one player, resulting in 1,707,997,180 video frames ($\sim$ 1.8 billion time steps) and 7907.4 hours of human gameplay. The training took 6 days on 16 GPUs (V100s) and the model was trained for $\sim$ 72000 steps (details in A.3.1).

AI Rollouts: We generated 600 rollouts of 1 min each with the above model by first picking a number of random game situations from the dataset. These are then filtered down to the desired characters e.g., Daemon, ZeroCool, and Makutu. (see appendix A.6 for details).

4 Behavioral analysis of gameplay data

4.1 Task-sets

We introduce the Task-sets framework for an interpretable analysis of human behavior from 100,000 games of Bleeding Edge (Power Collection game mode). We here formalize the definition of task-sets in conjunction with affordance and completion conditions over features of the (latent) state.

Definition 1 (Task-Set)**.** *A task-set comprises of extracting a set of features from the game state at each time step on which affordance and completion conditions are determined. Affordance conditions identify when a task-set can be performed or executed, giving the agent the choice of whether or not to engage in the task-set. Completion conditions determine if an agent successfully performed a task-set by choosing to engage. A task-set is said to be afforded when its affordance condition is met, and said to be completed when it's completion condition is met.*

This compositional approach to agent behavior, expressed as the composition of various task sets, underscores the flexibility and adaptability of the task-set framework in capturing agent behavior. To

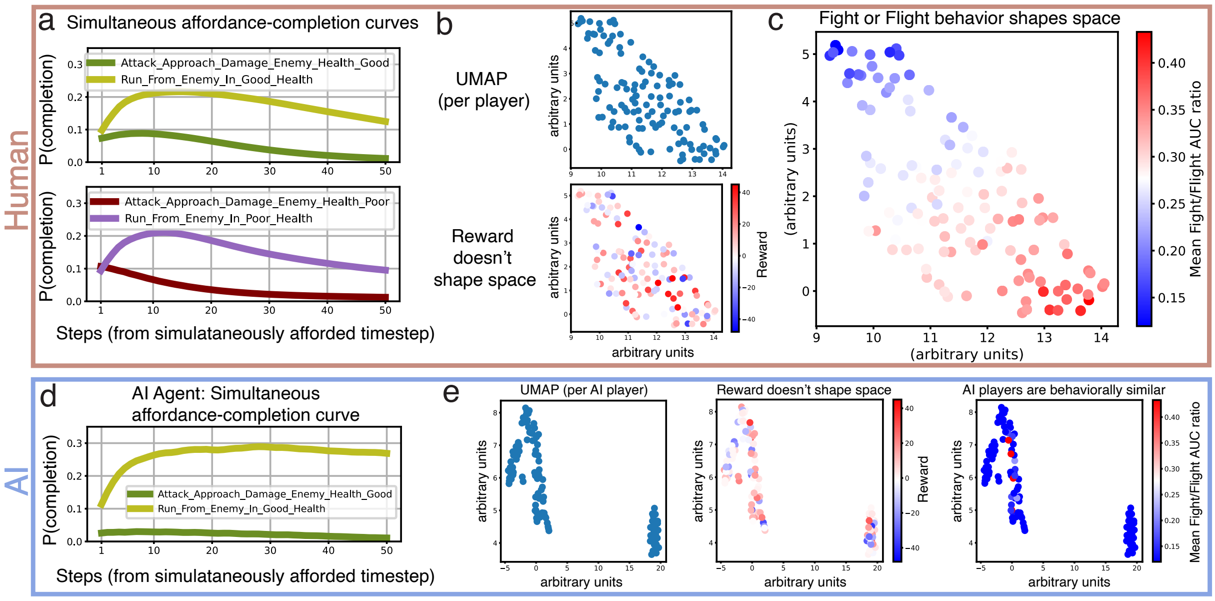


Figure 4: **Fight-Flight analysis results.** (a) Simultaneous affordance-completion curves for two representative pairs of fight-flight task-sets from human gameplay data. (b) *Top:* An unsupervised 2D UMAP (Uniform Manifold Approximation and Projection) embedding of 123 human players averaged across 637 games. Each point represents one human player. *Bottom:* Human UMAP colored by reward. (c) Human UMAP colored by fight-flight behavior. (d) Simultaneous affordance-completion curves for a representative pair of fight-flight task-set from AI gameplay data. (e) *Left:* UMAP embeding of 116 AI players averaged across 116 games. Each point represents one AI player. *Middle:* AI UMAP colored by reward. *Right:* AI UMAP colored by fight-flight behavior.

illustrate, consider the following task-sets (see the full list of all task-sets in Fig.A.16):

Run_From_Enemy_In_Good_Health
Affordance condition: the nearest enemy is (a) within 2100 distance units of our character, and (b) has above 50% of their health remaining.
Completion condition: our character is (a) moving away from the nearest enemy, and (b) that nearest enemy is within 3500 distance units.

Attack_Approach_Damage_Enemy_Health_Good
Affordance condition: the closest enemy to our character is (a) within 2100 distance units from the ego character, (b) they have over 50% of their health remaining, and (c) the ego character is moving toward them.
Completion condition: the ego character either dealt damage or was credited with a kill on this timestep.

We note that future work could consider using automatically learned task-sets similar to skill learning (Wang et al., 2023a; Khetarpal et al., 2021). However, in the scope of this work, we adhere to the programmer-specified definition of task sets based on our understanding of the game and the analysis of game play data, a deliberate choice aimed at illustrating the advantages inherent in this framework without being limited by the quality of learned task sets.

Cognitive themes: In our analytical framework, we employ three distinct cognitive themes that are ubiquitously observed in the behaviors of various biological species (Fig.3a). They are: 1)“Fight-Flight" for shedding light on the decision-making processes associated with confrontation and evasion, 2) “Explore-Exploit" is employed to discern the strategic balance between exploration and exploitation strategies within the game environment, and 3)“Solo-Multi-agent play" is used for understanding the interplay between individual and collaborative player behaviors. The task-sets defined for each of these cognitive themes are shown in Fig.3b (complete definitions in A.4).

4.2 Simultaneous affordance-completion analysis

We systematically analyze agent behavior across three aforementioned different axes that highlight meaningful variation: fight-flight, explore-exploit, and solo- vs. multi-agent play. For each axis, we identified a collection of task-sets that capture behavior along this axis, and conduct the following analysis:

Identification of Afforded Timesteps: For each

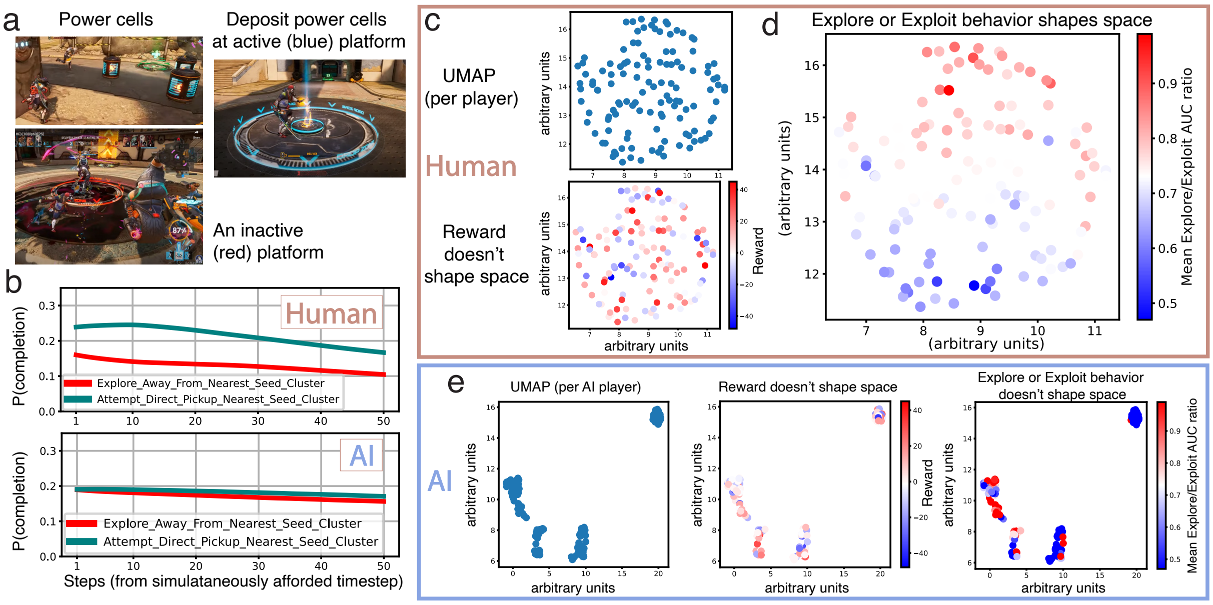


Figure 5: **Explore-Exploit analysis results.** (a) Goal-directed navigation based task-set illustrations. (b) Simultaneous affordance-completion curves for a representative pair of explore-exploit task-set from human *top* and AI (*bottom*) gameplay data. (c) *Top:* An unsupervised 2D UMAP embedding of 123 human players averaged across 637 games. Each point represents one human player. *Bottom:* Human UMAP colored by reward. (d) Human UMAP colored by explore-exploit behavior. (e) *Left:* UMAP embeding of 116 AI players averaged across 116 games. Each point represents one AI player. *Middle:* AI UMAP colored by reward. *Right:* AI UMAP colored by explore-exploit behavior.

timestep of every game in our dataset, we identify states at which all relevant task-sets were afforded to the agent.

Evaluation of Task-Set Completion: For each of the simultaneously afforded task-sets, we examine whether it was completed before it was afforded again to the agent. If completed, we record all future timesteps of potential completions before the next affordance.

Probability of Completion Computation: We aggregate this completion data across all simultaneous completion timesteps, and use it to compute the probability of completion of each of the afforded task-sets. Specifically, for each simultaneously afforded task-set, we compute the probability of completion at time $t + x$, given a simultaneous affordance at time t:

$$P(\text{completion at } t + x | \text{ simul. afford. at } t) = \frac{\text{\# completions of current task-set } x \text{ steps after affordance}}{\text{total \# of timesteps the task-sets were simultaneously afforded}}$$

That is, for each future step x, we add up the completion counts from different time-steps in which the given task-set combination was afforded, and divide by the total number of observations of the combination. Plotting this data produces the simultaneous affordance-completion curves for the task-sets (e.g., Fig.3c).

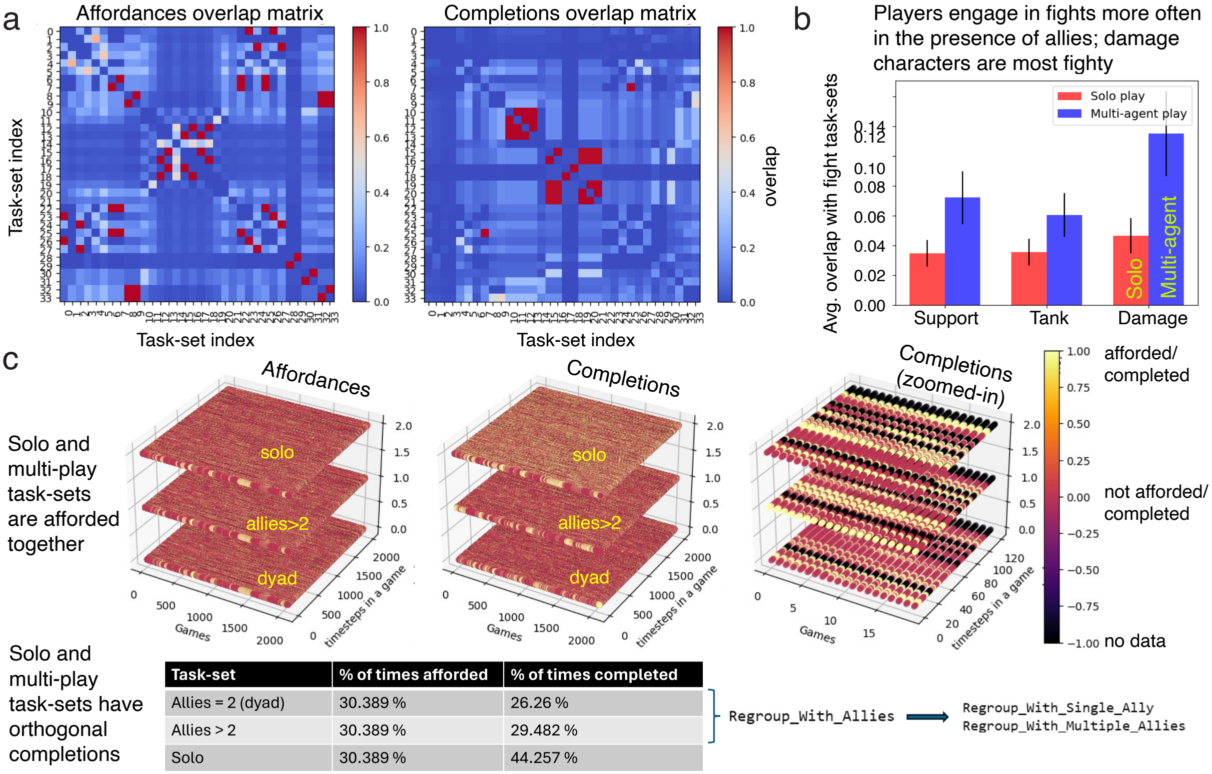
5 Results

5.1 Fight-Flight

Hypothesis: Players vary in how they express the fight or flight behavior while playing the same character.

To test this hypothesis, four pairs of task-sets (Fig.3b, right) that examine fight (attacking) or flight (running away) behavior were used for the analysis. Each pair has a different affordance criteria (appendix A.4.1). For each set of simultaneous affordances, we compute completion probabilities for both task-sets, and compare their completion probability curves as shown in Fig.4a. Note that the flight task-sets have a higher completion probability than the fight task-sets. For human gameplay data, we limited the analysis to the players who had played 3+ games controlling the Daemon character. Hence, the analysis was conducted for 123 players who played a total of 637 games with Daemon.

For each pair of task-sets, we generate 9 features: Area under the curve (AUC), Max, and Argmax of 'fight' task-set curve; AUC, Max, and Argmax of 'flight' task-set curve; and the ratio between each of the features, dividing fight / flight, resulting in



Task-set	% of times afforded	% of times completed
Allies = 2 (dyad)	30.389 %	26.26 %
Allies > 2	30.389 %	29.482 %
Solo	30.389 %	44.257 %

Figure 6: **Differences in human play-style across character types.** (a) *Left:* Average task-set affordances (*left*) and completions (*right*) overlap matrix for Damage characters. (b) Overlap of solo-multi-agent task-sets with fight task-sets averaged across all characters within each character type (Fig.1c). (c) *Left, Mid:* Affordances and completions respectively of solo-multi-agent task-sets (Fig.3b, Left) plotted for a subset of timesteps: ∼ 2000 time steps per game for 2059 games played by the Daemon character. *Right:* Zoomed in version of the completions plot showing orthogonal completions across the three task sets. *Bottom:* Table listing the average % of times each of the solo-multi-agent task sets were afforded/completed across all the characters and games.

$4 \times 9 = 36$ interpretable features per player. We use these features to produce an unsupervised 2D embedding of each human player through UMAP (Uniform Manifold Approximation and Projection) (McInnes et al., 2018) as shown in Fig.4b, top. Here each point corresponds to one human player leading to a total of 123 points on the manifold. The data for each player is averaged across all the games (3 or more) played by this player. When the manifold is colored by the reward (score) received by the players, we find that the reward doesn't shape the embedding space as shown in Fig.4b, bottom. However, when the manifold is colored by fight/flight AUC ratio, we find that the human behavioral manifold embedding space is shaped by fight or flight behavior as shown in Fig.4c. Although on an absolute scale, all players are flighty, relatively some players are more fighty (red points) than others.

With the same analysis on the AI agent gameplay data: Fig. 4d shows a representative simultaneous affordance-completion curve showing that the AI has a very low completion probability for fight task-sets. Fig.4e, left shows the UMAP embedding of 116 AI players averaged across 116 games, while playing Daemon. Here each point corresponds to one AI player leading to a total of 116 points on the manifold. Fig.4e, middle shows the manifold colored by reward received, indicating that the reward doesn't shape the AI embedding space. Fig.4e, right shows that when colored by fight/flight AUC ratio, almost all AI players are behaviorally similar, i.e., almost all of them are flighty. This is in contrast to the human players who are behaviorally different on a relative scale with some of them being more fighty and some being more flighty (Fig.4c).

Conclusion: We find that while human players vary in how they express the fight or flight behavior while playing the same character, AI players do not. Apart from Daemon, we also ran the above analysis for other characters in Fig.1c and drew the same conclusion irrespective of the character.

5.2 Explore-Exploit

Hypothesis: Players vary in how they express the explore or exploit behavior while playing the same character.

To test this hypothesis we defined three pairs of goal-directed navigation based task-sets listed in Fig.3b, middle and shown in Fig.5a. Exploitation prioritizes immediate rewards by efficiently achieving the objectives in a direct and goal-oriented manner. Similar to the fight-flight analysis above (section 5.1), we compute the simultaneous affordance-completion curves for each pair of explore-exploit task-sets for human players as well as AI agents. Fig.5b,top shows that humans are exploiters - the exploit task-set has a higher completion probability than the explore task-set. Next, we compute the UMAP embedding as shown in Fig.5c,top. Again, we find that the reward doesn't shape human behavioral manifold embedding space (Fig.5c, bottom), however, when colored by explore/exploit AUC ratio, the embedding space is shaped by explore or exploit behavior (Fig.5d). Although on an absolute scale, all players are exploiters, relatively some players explore more (red points) than others.

We ran the same analysis on the AI agent gameplay data. Fig. 5b,bottom shows that the AI has almost equal completion probability for the explore and exploit task-set. Fig. 4e,left shows the UMAP embedding of AI players, while playing Daemon. Fig.4e,middle shows the manifold colored by reward received, indicating that the reward doesn't shape the AI embedding space. Fig.4e,right shows that when colored by explore/exploit AUC ratio, the AI embedding space is not shaped by explore-exploit behavior either. This indicates that although AI players show some diversity in exploring vs exploiting, both of these groups share similar feature representations, making it difficult to differentiate between them based on explore-exploit behavior. In other words, AI agents exhibit overlapping characteristics between exploration and exploitation behavior and can't be grouped based on it. This is in contrast to the human players who are behaviorally different on a relative scale with some of them being more of explorers and some being more of exploiters (Fig.5d).

Conclusion: We find that while human players clearly vary in how they express explore or exploit behavior when playing the same character, we can't differentiate between AI players based on this behavior. This conclusion stands across characters.[1]

5.3 Solo-Multi-agent play

To identify differences in play-style across character types, we compute the overlap in task-set affordances and completions (Fig.6a) for each character type (Fig.A.8a, A.11a). On analysing the difference between these overlap matrices (Fig.A.8b, A.11b), we find that the Tank characters are the most suited to carrying power cells (Fig.A.10), Support characters to healing (Fig.A.12), and Damage characters to dealing damage (Fig.A.13). From the completions overlap matrices, we extract the overlap of completions of solo-multi-agent game play task sets (Fig.6b, left) with the fight task-sets in the human gameplay data. We find that irrespective of the character type, all human players engage in fights more often in the presence of allies relative to when they are playing solo. Additionally, we find that the Damage characters are the most fighty.

Besides, we find that all the solo-mutli-agent task-sets are afforded simultaneously as is evident by the three identical affordance planes shown in Fig.6c,left. However, all of them have orthogonal completions illustrated by the three distinct completion planes (see Fig.6c, middle and right). The table in the figure summarizes the fraction of times the three task-sets are afforded and completed (relative to affordances) in the game. Although all of them are simultaneously afforded, their completions are orthogonal, leading to completion fractions that sum up to 100 for the three task sets. When playing the Daemon character, humans play in cooperation with their allies $\approx 56\%$ of the time, and play solo only $\approx 44\%$ of the time.

Table 2 (in appendix) lists the percentage (%) of time spent in solo vs multi-agent game play by human players when playing different characters (Fig.1c). Irrespective of the character played, humans spend a significantly greater amount of time playing cooperatively with their allies in multi-agent settings relative to solo game play. We subject the AI game play data to the same analysis for a single randomly chosen character for each character type with results summarized in Table 3. We find that in contrast to humans, AI agents spend majority of game time ($\approx 70\%$) playing solo.

[1] We find that the same analysis for different characters in Fig.1c results in the same findings.

6 Discussion

We propose a Task-sets-based framework for (i) understanding human behavior in large-scale multiplayer games, and (ii) assessing the alignment of AI agents with humans. We apply this framework to examine the alignment of a proof of concept GPT-based AI agent trained to play Bleeding Edge. Through simultaneous affordance-completion analysis of task-sets, we examine interpretable behavioral axes, allowing for richer comparisons than what policy differences alone allow for.

Our analysis of the human data shows that the task-sets capture meaningful factors of variation in human behavior along the three behavioral axes. While humans show substantial differences along these dimensions, data from our AI agent does not mirror the variations observed in human behavior. We take this as evidence that these AI agents are not aligned with humans. See A.7 for future work, broader impact and societal implications.

References

Andrej Karpathy. 2023. The simplest, fastest repository for training/finetuning medium-sized GPTs.

AutoGPT. 2023. An experimental open-source attempt to make gpt-4 fully autonomous.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*.

Pierre-Luc Bacon, Jean Harb, and Doina Precup. 2017. The option-critic architecture. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. 2022. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654.

Andreea Bobu, Andi Peng, Pulkit Agrawal, Julie Shah, and Anca D Dragan. 2023. Aligning robot and human representations. *arXiv preprint arXiv:2302.01928*.

Erik Brynjolfsson, Danielle Li, and Lindsey R Raymond. 2023. Generative ai at work. Technical report, National Bureau of Economic Research.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*.

Tathagata Chakraborti and Subbarao Kambhampati. 2018. Algorithms for the greater good! on mental modeling and acceptable symbiosis in human-ai collaboration. *arXiv preprint arXiv:1801.09854*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Anne GE Collins and Michael J Frank. 2013. Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological review*, 120(1):190.

Yuqing Du, Olivia Watkins, Zihan Wang, Cédric Colas, Trevor Darrell, Pieter Abbeel, Abhishek Gupta, and Jacob Andreas. 2023. Guiding pretraining in reinforcement learning with large language models. *arXiv preprint arXiv:2302.06692*.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437.

Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pretrained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Krishnaram Kenthapadi, Himabindu Lakkaraju, and Nazneen Rajani. 2023. Generative ai meets responsible ai: Practical challenges and opportunities. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5805–5806.

Khimya Khetarpal, Zafarali Ahmed, Gheorghe Comanici, and Doina Precup. 2021. Temporally abstract partial models. *Advances in Neural Information Processing Systems*, 34:1979–1991.

Khimya Khetarpal, Martin Klissarov, Maxime Chevalier-Boisvert, Pierre-Luc Bacon, and Doina Precup. 2020. Options of interest: Temporal abstraction with interest functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4444–4451.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. 2022. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*.

Yoonho Lee, Annie S Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. 2022. Surgical fine-tuning improves adaptation to distribution shifts. *arXiv preprint arXiv:2210.11466*.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Ida Momennejad and John-Dylan Haynes. 2013. Encoding of prospective tasks in the human prefrontal cortex under varying task loads. *Journal of Neuroscience*, 33(44):17342–17349.

Stephen Monsell. 2003. Task switching. *Trends in cognitive sciences*, 7(3):134–140.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Dean A Pomerleau. 1991. Efficient training of artificial neural networks for autonomous navigation. *Neural computation*, 3(1):88–97.

Doina Precup. 2000. *Temporal abstraction in reinforcement learning*. University of Massachusetts Amherst.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Katsuyuki Sakai. 2008. Task set and prefrontal cortex. *Annu. Rev. Neurosci.*, 31:219–245.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Sugandha Sharma, Aidan Curtis, Marta Kryven, Joshua B. Tenenbaum, and Ila R Fiete. 2022. Map induction: Compositional spatial submap learning for efficient exploration in novel environments. In *International Conference on Learning Representations*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. 2014. Deterministic policy gradient algorithms. In *International conference on machine learning*, pages 387–395. Pmlr.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Martin Stolle and Doina Precup. 2002. Learning options in reinforcement learning. In *Abstraction, Reformulation, and Approximation: 5th International Symposium, SARA 2002 Kananaskis, Alberta, Canada August 2–4, 2002 Proceedings 5*, pages 212–223. Springer.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Jascha Achterberg, Joshua B Tenenbaum, et al. 2023. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999a. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12.

Richard S Sutton, Doina Precup, and Satinder Singh. 1999b. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211.

Avinash R Vaidya and David Badre. 2022. Abstract task representations for inference and control. *Trends in Cognitive Sciences*.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023b. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

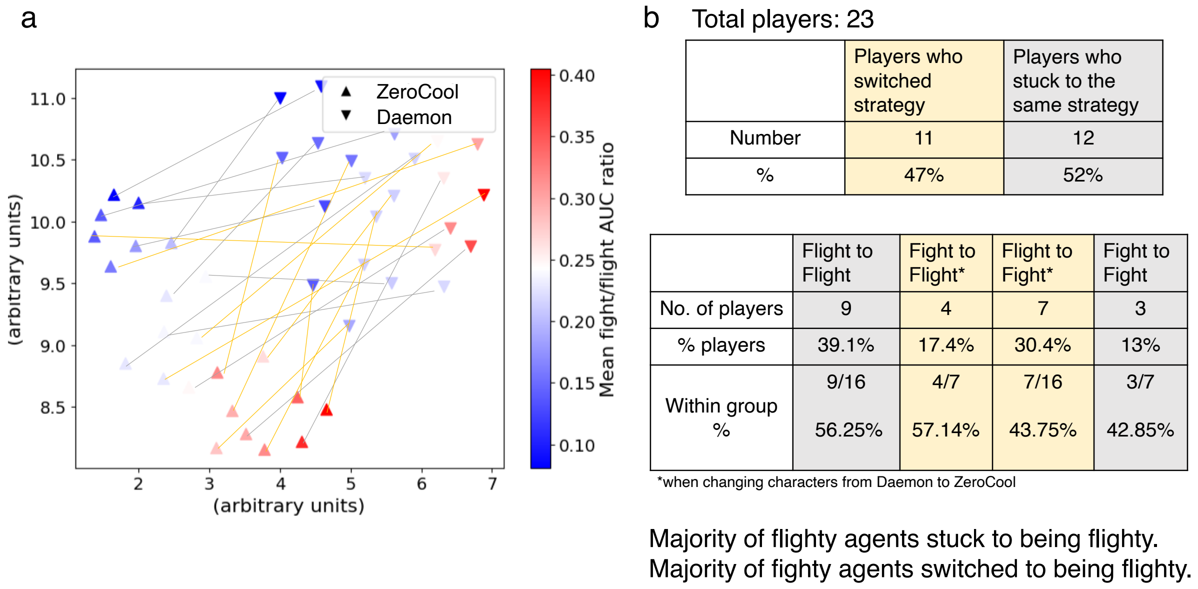
Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. 2022. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971.

Andrea Wynn, Ilia Sucholutsky, and Thomas L Griffiths. 2023. Learning human-like representations to enable learning human values. *arXiv preprint arXiv:2312.14106*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

A Appendix



Total players: 23

	Players who switched strategy	Players who stuck to the same strategy
Number	11	12
%	47%	52%

	Flight to Flight	Fight to Flight*	Flight to Fight*	Fight to Fight
No. of players	9	4	7	3
% players	39.1%	17.4%	30.4%	13%
Within group %	9/16 56.25%	4/7 57.14%	7/16 43.75%	3/7 42.85%

*when changing characters from Daemon to ZeroCool

Majority of flighty agents stuck to being flighty.
Majority of fighty agents switched to being flighty.

Figure A.7: **Change in player strategies when changing characters.** *Left:* A manifold showing the change in strategies when switching between ZeroCool and Daemon characters. *Right:* Summary of results from the switching dynamics analysis of all players who played 3+ games with both ZeroCool and Daemon. Out of the players who switched strategies, a greater % of players switched to Flight relative to switching to Fight, and out of the players who stuck to the same strategy, a greater % were flighty rather than fighty.

A.1 Acknowledgements

We would like to acknowledge Tabish Rashid at Microsoft Research for support and guidance on the AI agent. We would also like to acknowledge the Ninja Theory team (that developed Bleeding Edge) for their feedback during the course of this project.

A.2 Why Bleeding Edge

Although there are multiple games that we could have chosen for the work presented in this paper, we opted to work with Bleeding Edge since we had access to rich human behavioral data from human game play in this game that was easy to work with and process, made possible by our current affiliations. This data was used to train the AI agents presented in the paper as well as for human behavioral analysis through the task-sets framework in order to study human-AI alignment.

A.3 AI Agent details

A.3.1 Training

For this work, we train the agent on less than one epoch for computational time and memory complexity reasons. While training we add data augmentation to the video frames following Baker et al. (2022).

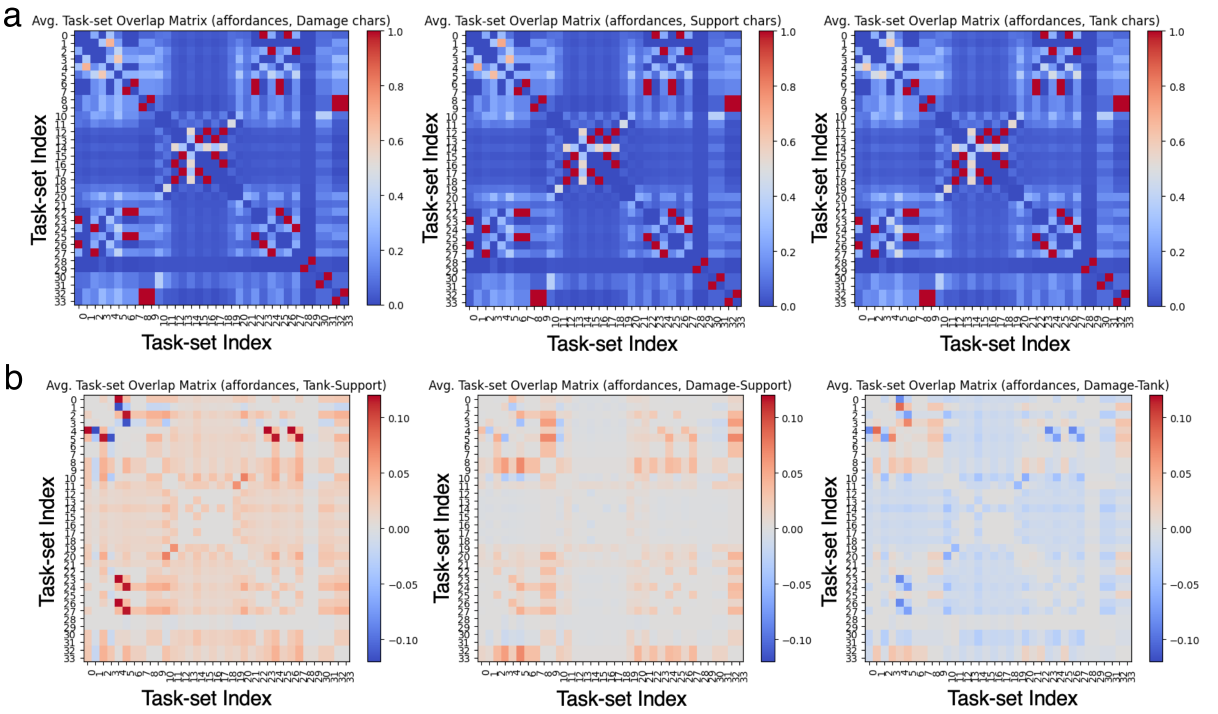


Figure A.8: **Overlap in task-set affordances.** (a) Avg. task-set affordance overlap matrix for *Left:* Damage characters, *Middle:* Support characters, *Right:* Tank characters. (b) Difference in the matrices in (a). *Left:* Tank-Support, *Left:* Damage-Support, *Left:* Damage-Tank.

Table 1: Hyperparameters for AI agent training

Parameter	**Value**
Steps	72,000
Learning Rate	0.0001
Warmup Steps	1000
Optimizer	AdamW
Optimizer weight decay	0.0001
Batch Size	12
Sequence length	128
L2 Gradient Clipping	1.0

The learning rate is scheduled by the following function:

$$lr = min((steps + 1)/warmupSteps, 1) \tag{1}$$

A.3.2 Observation Encoder

The model is trained on sequences of $T = 128$ images where each image is reshaped to $128 \times 128 \times 3$ and then divided by 255 to ensure its value lies in range [0, 1]. A custom ResNet (He et al., 2016) with 18.6M parameters is used to embed each image observation ($O \in R^{3\times128\times128}$) independently into a vector. The first layer is a 2D convolutional network with kernels of shape 5×5, a stride of 3, and a padding of 1 and maps to 64 channel dimension. We then apply GELU (Hendrycks and Gimpel, 2016) activation. This is followed by 4 ConvNext (Liu et al., 2022) and downsampling blocks. Each downsampling block applies group normalization and a convolution layer with kernel of shape 3×3, stride of 2, and padding of 1, doubling the number of channels. We again apply GELU activation followed by another 2D convolutional network with a kernel of shape 3×3, stride of 1×3 and padding of 0. For each input image, the output of the encoder is a 1024D embedding.

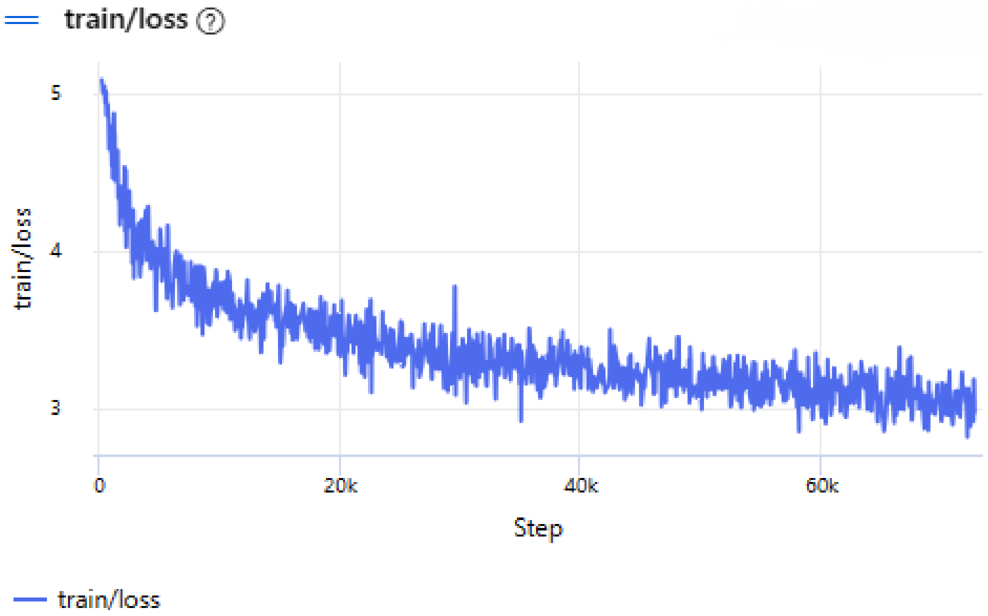


Figure A.9: **AI agent loss curve**. The decreasing loss curve indicates that over time, the AI agent improves its performance and makes action predictions that are increasingly closer to the ground truth. The smooth, steady decrease in the loss suggests that the AI agent is learning effectively and converging towards an optimal solution.

A.4 Task-set definitions

A.4.1 Fight-Flight

We define four pairs of task-sets, one each for fight and flight:

1. Absolute enemy health $> 50\%$:
   Affordance condition: the nearest enemy is within 2100 distance units from the ego character, has above 50% ('good') of their health remaining and the ego character is moving toward them.

   (a) *Fight: Attack_Approach_Damage_Enemy_Health_Good*
   Completion condition: the ego character dealt damage on this timestep.

   (b) *Flight: Run_From_Enemy_In_Good_Health*
   Completion condition: the ego character is moving away from the nearest enemy, and the nearest enemy is within 3500 distance units from the ego character.

2. Absolute enemy health $< 50\%$:
   Affordance condition: the nearest enemy is within 2100 distance units from the ego character, and has below 50% ('poor') of their health remaining.

   (a) *Fight: Attack_Approach_Damage_Enemy_Health_Poor*
   Completion condition: the ego character dealt damage on this timestep.

   (b) *Flight: Run_From_Enemy_In_Poor_Health*
   Completion condition: the ego character is moving away from the nearest enemy, and the nearest enemy is within 3500 distance units from the ego character.

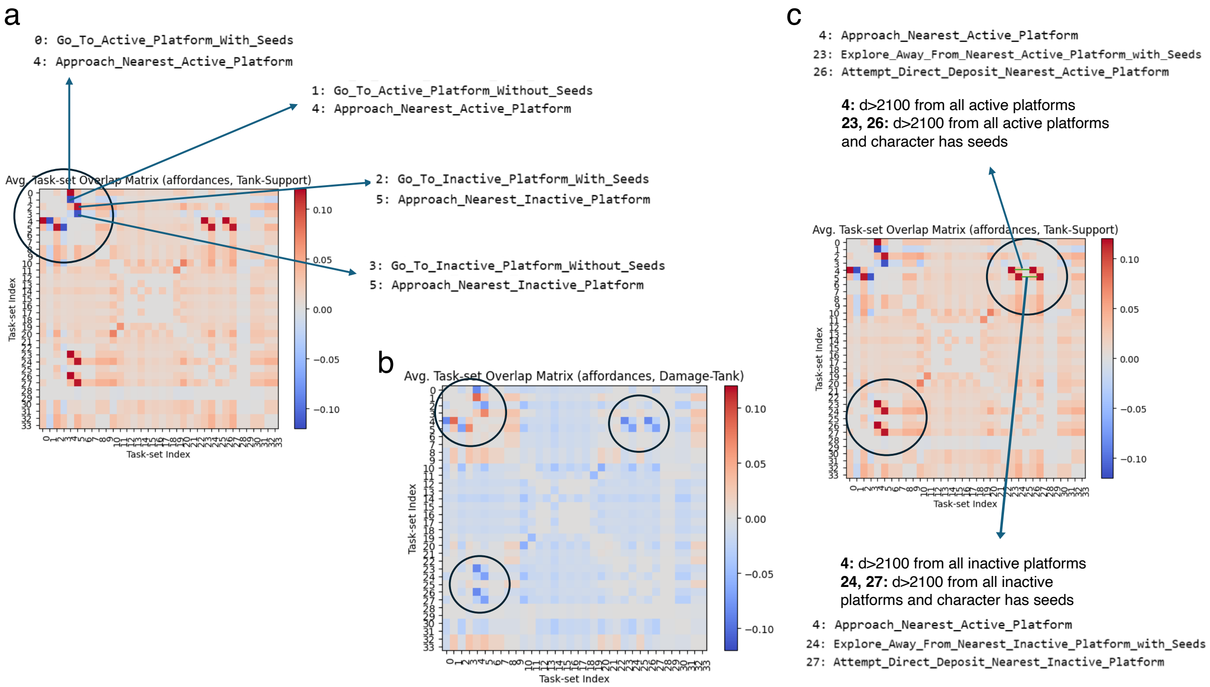


Figure A.10: **Analysis of overlap in task-set affordances shows that Tanks are the most suited to carrying power cells.** (a, c) Tanks have a higher overlap with task-sets that involve going to a platform with power cells showing that Tanks are more likely to go to platform with seeds than Supports. (b) Tanks are more likely to go to a platform with seeds than Damage characters.

3. Enemy health $>$ Player health:
   Affordance condition: the nearest enemy is within 2100 distance units, they have a larger ('greater') % of their health remaining than the ego character, and the ego character took damage on this timestep.

   (a) *Fight: Fight_Damage_Enemy_When_Attacked_Enemy_Health_Greater*
   Completion condition: the ego character dealt damage on this timestep.

   (b) *Flight: Run_When_Attacked_Enemy_Health_Greater*
   Completion condition: the ego character is moving away from the nearest enemy, and the nearest enemy is within 3500 distance units from the ego character.

4. Enemy health $<$ Player health:
   Affordance condition: the nearest enemy is within 2100 distance units, they have a lower ('poorer') % of their health remaining than the ego character, and the ego character took damage on this timestep.

   (a) *Fight: Fight_Damage_Enemy_When_Attacked_Enemy_Health_Poorer*
   Completion condition: the ego character dealt damage on this timestep.

   (b) *Flight: Run_When_Attacked_Enemy_Health_Poorer*
   Completion condition: the ego character is moving away from the nearest enemy, and the nearest enemy is within 3500 distance units from the ego character.

A.4.2 Explore-Exploit

We define three pairs of task-sets, one each for explore and exploit:

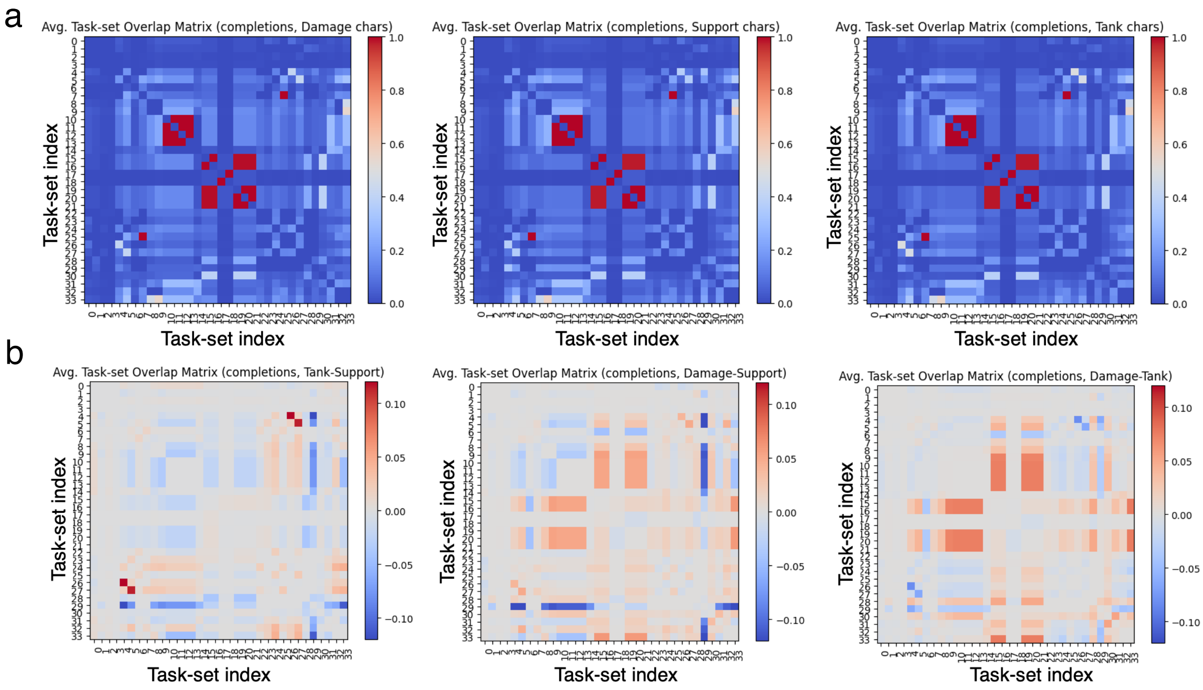


Figure A.11: **Overlap in task-set completions.** (a) Avg. task-set completions overlap matrix for *Left:* Damage characters, *Middle:* Support characters, *Right:* Tank characters. (b) Difference in the matrices in (a). *Left:* Tank-Support, *Left:* Damage-Support, *Left:* Damage-Tank.

1. Seed collection strategy:
   Affordance condition: there exists at least one visible seed cluster, ego character distance $>$ 2100 from all visible seed clusters.

   (a) *Exploit: Attempt_Direct_Pickup_Nearest_Seed_Cluster*
   Completion condition: there exists at least one visible seed cluster, the ego character is moving towards the nearest seed cluster.

   (b) *Explore: Explore_Away_From_Nearest_Seed_Cluster*
   Completion condition: the ego character is moving away from the nearest seed cluster.

2. Deposit strategy (relevant in deposit phase):
   Affordance condition: the ego character has seeds (number of seeds $>$ 0) and distance of ego character $>$ 2100 units from all active platforms .

   (a) *Exploit: Attempt_Direct_Deposit_Nearest_Active_Platform*
   Completion condition: the ego character has seeds (number of seeds $>$ 0), and is moving towards the nearest active platform.

   (b) *Explore: Explore_Away_From_Nearest_Active_Platform_with_Seeds*
   Completion condition: the ego character is moving away from the nearest active platform.

3. Deposit strategy (relevant in collection phase):
   Affordance condition: the ego character has seeds (number of seeds $>$ 0), ego character distance

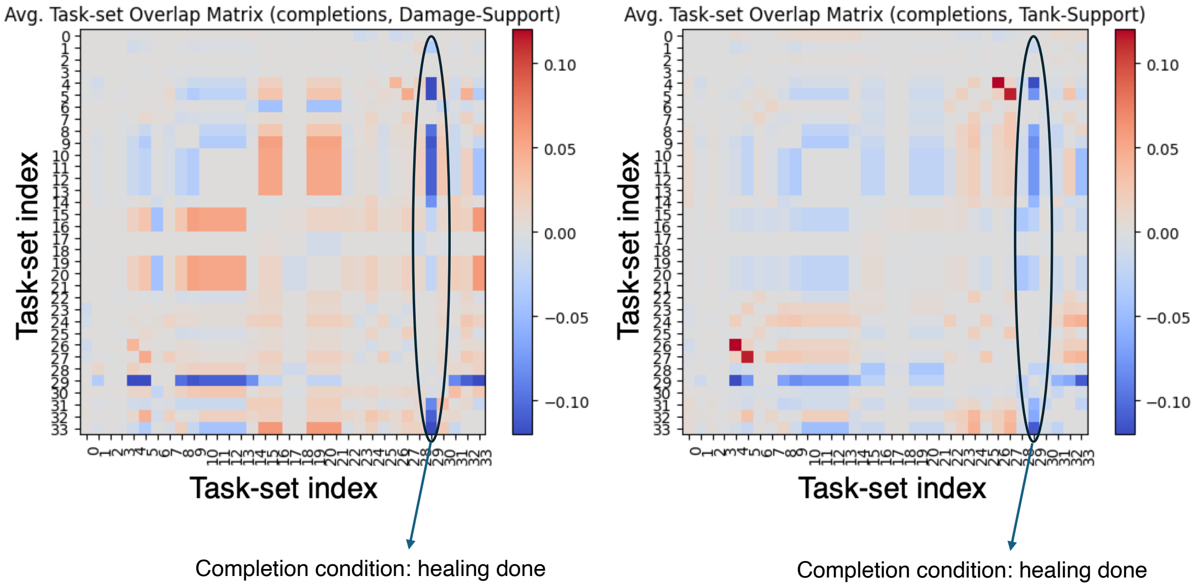


Figure A.12: **Analysis of overlap in task-set completions shows that Supports are the most suited to healing.** Difference in completions matrices of Damage and Support as well as Tank and Support characters shows that Supports have a higher overlap with all the task-sets that have a completion condition that includes healing.

$>$ 2100 units from all inactive platforms.

(a) *Exploit: Attempt_Direct_Deposit_Nearest_Inactive_Platform*
Completion condition: the ego character has seeds (number of seeds > 0), and is moving towards the nearest inactive platform.

(b) *Explore: Explore_Away_From_Nearest_Inactive_Platform_with_Seeds*
Completion condition: the ego character is moving away from the nearest inactive platform.

A.4.3 Solo-Multi

We define task-sets to study solo vs multi-agent game play dynamics.

Affordance condition: no single teammates within a distance of 3500

1. *Solo: Continue_To_Play_Solo*
Completion condition: distance from nearest teammate $>$ 2100.

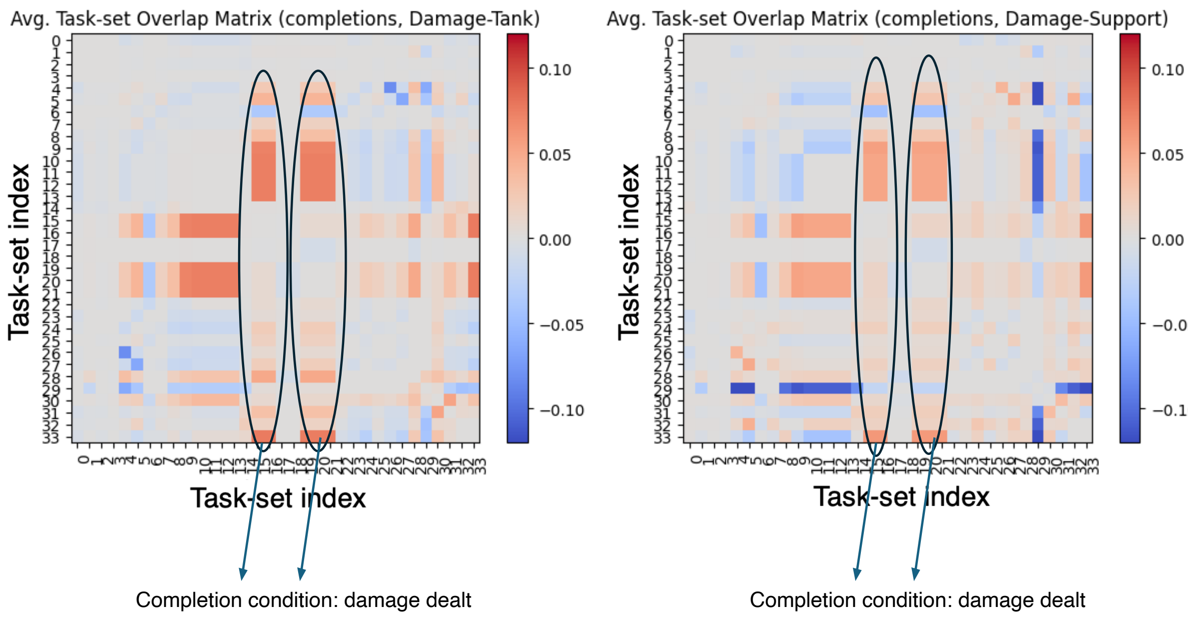
2. *Regroup: Regroup_With_Allies*
Completion condition: distance from nearest teammate $<$ 2100.

3. *Diad: Regroup_With_Single_Ally*
Completion condition: distance from nearest teammate $<$ 2100 and only one teammate is present within this distance range.

4. *Multi-agent: Regroup_With_Multiple_Allies*
Completion condition: distance from multiple (more than one) teammates $<$ 2100.

A.5 Character types

When playing Bleeding Edge, players select their character, from a diverse roster of 13 characters, each with a unique set of abilities and playstyles. The characters are classified into three main categories (Fig.1c):



Completion condition: damage dealt

Completion condition: damage dealt

Figure A.13: **Analysis of overlap in task-set completions shows that Damage characters are the most suited to dealing damage.** Difference in completions matrices of Damage and Tank as well as Damage and Support characters shows that Damage characters have a higher overlap with all the task-sets that have a completion condition that includes dealing damage.

1. Support: Possess healing abilities, buffs, crowd control, or other utility tools. They excel at keeping their allies alive, providing additional damage or defense boosts, and disrupting the enemy team.
2. Tank: Durable and resilient, capable of soaking up large amounts of damage and protecting their teammates. They have high health pools and often possess abilities that allow them to mitigate or redirect damage away from their more fragile allies.
3. Damage: Excel at engaging in fights and eliminating opponents quickly. They tend to have lower health pools and may require support or protection from Tank characters to survive in prolonged engagements.

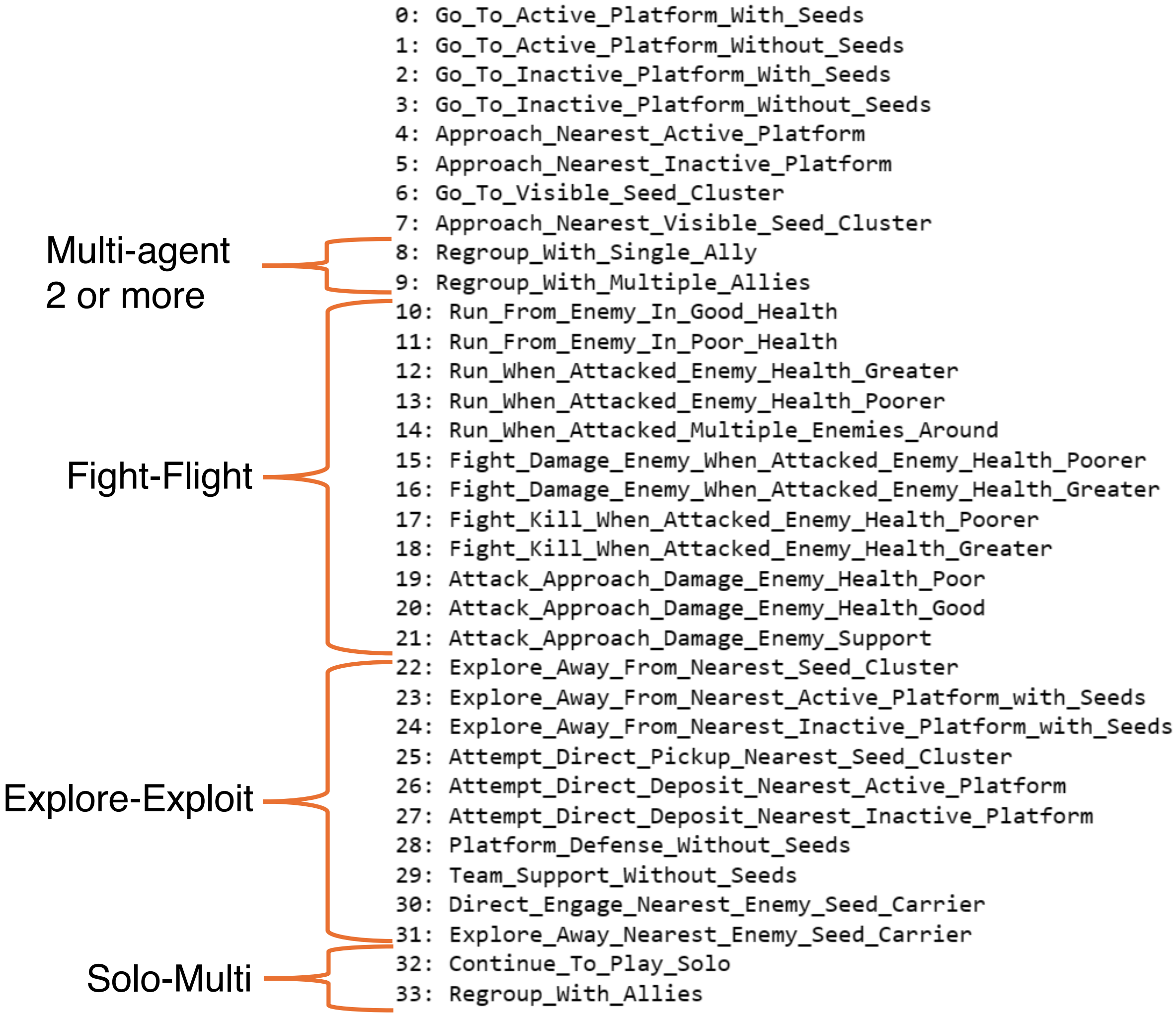
The varied selection of characters allows for strategic team composition, encouraging players to tailor their choices to complement their team's overall strategy. This diversity promotes collaborative and strategic thinking as players work together to capitalize on each character's strengths.

A.6 AI Rollouts

For each game, we initialize the game at that state and let the AI agent play from that state onwards for a duration of 1 minute. Except for the AI agent, all other players in the game are non-players who aren't controlled and simply continue to repeat the latest action that the corresponding human player acted during the real human gameplay (e.g., if the player was moving forward when the game was originally recorded at that state, they will continue moving forward). They do not respond to the AI agent's actions, however, their features get affected based on AI agent's actions. For instance, their health decreases when the AI agent attacks them. We assume that each rollout is a different AI player given the stochasticity e.g., due to sampling the action from the output probability distribution.

A.7 Future work, impact and implications

Future work: An important direction is to assess how the identified behavioral dimensions can be used practically to develop agents that demonstrate targeted behaviors. Our analyses identify different play styles, such as aggressive (fighty) behavior. Player data could be separated by play style and used to fine-tune different agents, which may then replicate this behavior more readily. A second promising



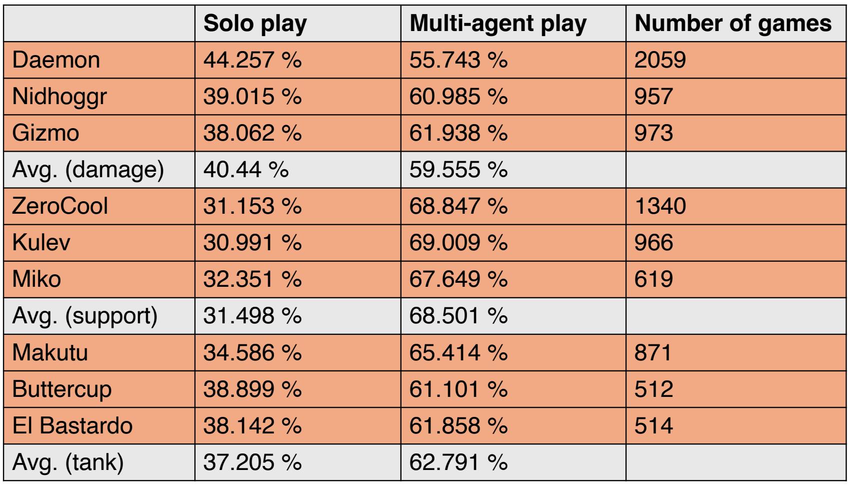
```
0: Go_To_Active_Platform_With_Seeds
1: Go_To_Active_Platform_Without_Seeds
2: Go_To_Inactive_Platform_With_Seeds
3: Go_To_Inactive_Platform_Without_Seeds
4: Approach_Nearest_Active_Platform
5: Approach_Nearest_Inactive_Platform
6: Go_To_Visible_Seed_Cluster
7: Approach_Nearest_Visible_Seed_Cluster
8: Regroup_With_Single_Ally
9: Regroup_With_Multiple_Allies
10: Run_From_Enemy_In_Good_Health
11: Run_From_Enemy_In_Poor_Health
12: Run_When_Attacked_Enemy_Health_Greater
13: Run_When_Attacked_Enemy_Health_Poorer
14: Run_When_Attacked_Multiple_Enemies_Around
15: Fight_Damage_Enemy_When_Attacked_Enemy_Health_Poorer
16: Fight_Damage_Enemy_When_Attacked_Enemy_Health_Greater
17: Fight_Kill_When_Attacked_Enemy_Health_Poorer
18: Fight_Kill_When_Attacked_Enemy_Health_Greater
19: Attack_Approach_Damage_Enemy_Health_Poor
20: Attack_Approach_Damage_Enemy_Health_Good
21: Attack_Approach_Damage_Enemy_Support
22: Explore_Away_From_Nearest_Seed_Cluster
23: Explore_Away_From_Nearest_Active_Platform_with_Seeds
24: Explore_Away_From_Nearest_Inactive_Platform_with_Seeds
25: Attempt_Direct_Pickup_Nearest_Seed_Cluster
26: Attempt_Direct_Deposit_Nearest_Active_Platform
27: Attempt_Direct_Deposit_Nearest_Inactive_Platform
28: Platform_Defense_Without_Seeds
29: Team_Support_Without_Seeds
30: Direct_Engage_Nearest_Enemy_Seed_Carrier
31: Explore_Away_Nearest_Enemy_Seed_Carrier
32: Continue_To_Play_Solo
33: Regroup_With_Allies
```

Figure A.14: **Task-sets**. List of all task-sets implemented for the analysis presented in this paper. Each of these was implemented as a routine in python

direction is extending the task-sets framework by incorporating automatic discovery and learning of task-sets. This would increase the framework's applicability to additionaldomains and make it easier to apply without a preceding substantial data analysis effort. Finally, a third direction could explore whether any model parameters or components in the latent representations are associable with specific axes of the behavioral manifold. This may offer mechanistic insights into the factors influencing high-level behavior, and therefore human-agent alignment.

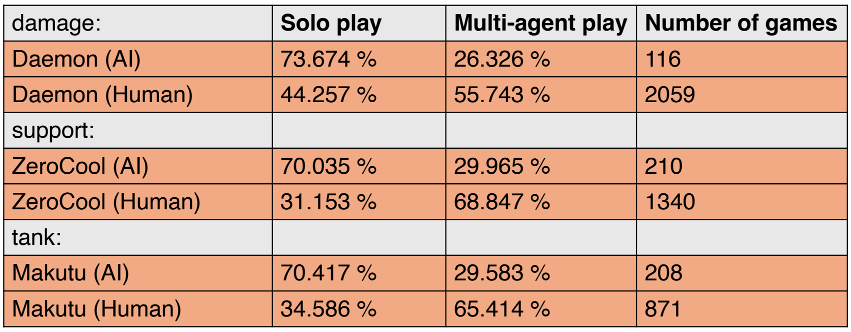
Broader impact: Our work shows that the alignment of transformer-based models trained on next token prediction may not always be inherent, and it may require specialized training techniques such as supervised fine-tuning (Gunel et al., 2020; Wortsman et al., 2022; Lee et al., 2022; Kirichenko et al., 2022) and Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Stiennon et al., 2020; Glaese et al., 2022; Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2023; Azar et al., 2023) to achieve alignment. This underscores the importance of our framework for measuring alignment in AI agents.

Broader societal implications: Aligning AI with humans along fight-flight responses can help address ethical and moral questions about the use of AI in simulated (and potentially real) conflict situations, defense systems, and decision-making processes that involve risk, uncertainty, and potential harm to individuals and communities. AI alignment with human preferences for solo or multiplayer gameplay can influence social interactions, cultural norms, and community dynamics in gaming and entertainment,



	Solo play	Multi-agent play	Number of games
Daemon	44.257 %	55.743 %	2059
Nidhoggr	39.015 %	60.985 %	957
Gizmo	38.062 %	61.938 %	973
Avg. (damage)	40.44 %	59.555 %	
ZeroCool	31.153 %	68.847 %	1340
Kulev	30.991 %	69.009 %	966
Miko	32.351 %	67.649 %	619
Avg. (support)	31.498 %	68.501 %	
Makutu	34.586 %	65.414 %	871
Buttercup	38.899 %	61.101 %	512
El Bastardo	38.142 %	61.858 %	514
Avg. (tank)	37.205 %	62.791 %	

Table 2: Percentage of time spent by human players playing solo (player playing alone) is substantially less than that spent playing with their allies (multi-agent: 2 allies or more playing together) for the characters in three character types (Fig.1c) averaged across all games played by the given character.



damage:	Solo play	Multi-agent play	Number of games
Daemon (AI)	73.674 %	26.326 %	116
Daemon (Human)	44.257 %	55.743 %	2059
support:			
ZeroCool (AI)	70.035 %	29.965 %	210
ZeroCool (Human)	31.153 %	68.847 %	1340
tank:			
Makutu (AI)	70.417 %	29.583 %	208
Makutu (Human)	34.586 %	65.414 %	871

Table 3: Percentage of time spent by AI vs Human players playing alone (solo) and with their allies (multi-agent) for one character from each of the three character types (Fig. 1c) averaged across all games played by a given character. AI predominantly plays solo.

shaping how individuals and groups engage with AI-driven gaming experiences, collaborate with AI agents, and form online (and potentially offline) communities. Human-AI alignment along exploration or exploitation can foster innovation, creativity, and adaptive learning in various domains, including research, development, entrepreneurship, and education.

Our framework can contribute to the long-term sustainability of AI technologies by promoting human-centered approaches to AI development and deployment. Since it is an interpretable framework, it can enhance user confidence in AI and promote their widespread adoption.

A.8 Ethics review program for human data collection

At Microsoft Research, we have an internal ethics review program that helped us work with human data in a way that ensured respect and protection of the rights of human participants contributing to our research.

Our institutional Review Board (IRB) approved all the data collection for the data used in this paper. For the player recordings, we received ethics approval (IRB 10601) from our organization's Ethics Review Program. Our organization's Institutional Review Board (IRB) has been officially registered with the U.S. Health and Human Services' Office for Human Research Protections (OHRP) since 2017.

During data collection, human players playing Bleeding Edge are subjected to a pop up when they log in to the game for the first time to agree to the terms (end-user license agreement). The behavioral data collected does not contain any personally identifiable information.

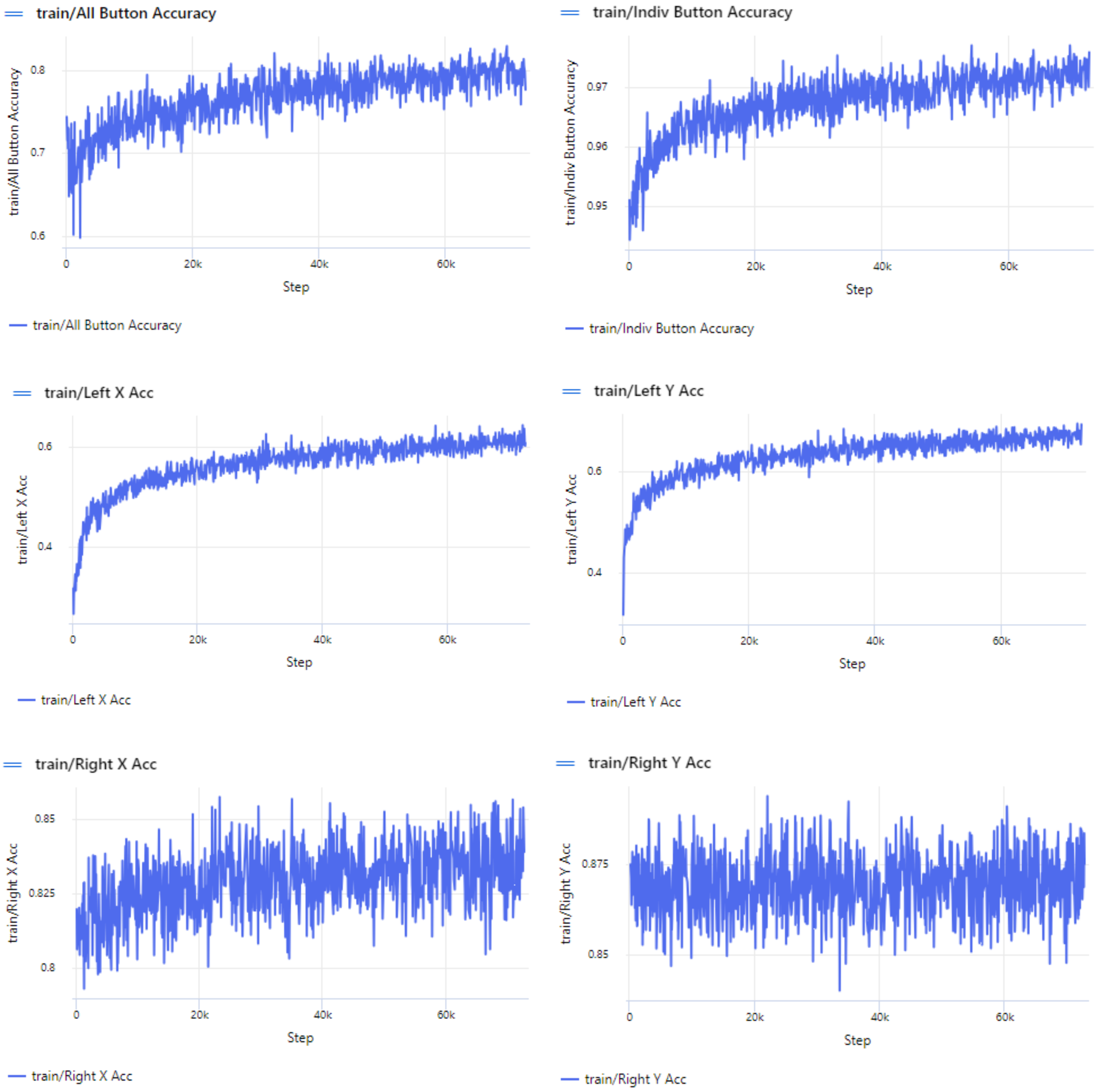


Figure A.15: **AI agent action accuracy curves over training**. The Xbox controller action space consists of 12 discrete buttons and two joysticks. *Top:* Button accuracy curves for all the buttons (percentage of times the predicted value of all buttons matches the expected value) on the left; and for individual buttons (mean percentage of times the predicted value of each individual button matches its expected value). *Middle:* Left joystick accuracy curves for x and y components of the joystick. This joystick controls character movement. *Bottom:* Right joystick accuracy curves for x and y components of the joystick. This joystick controls the camera.

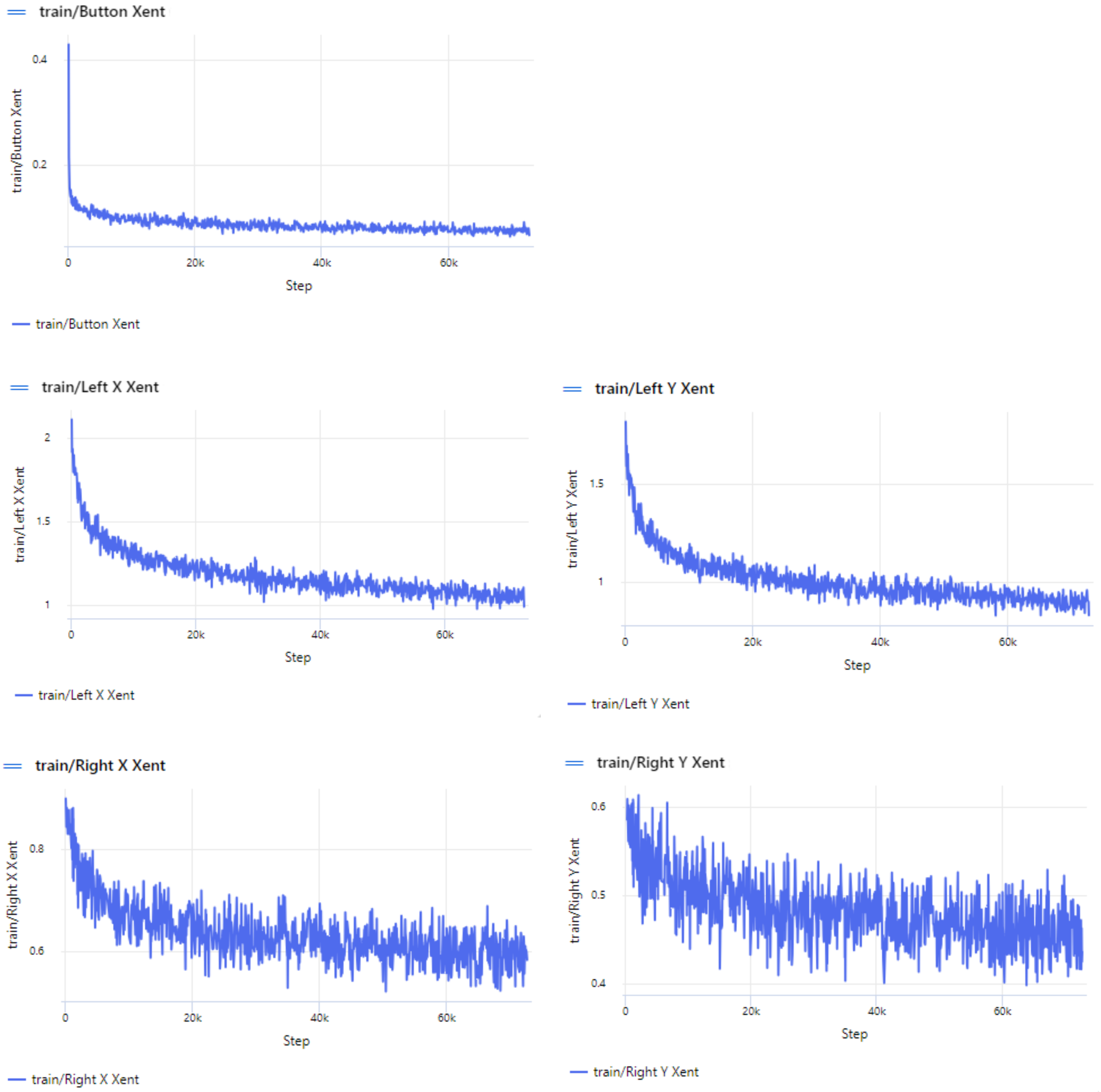


Figure A.16: **AI agent decomposed loss curves over training**. The Xbox controller action space consists of 12 discrete buttons and two joysticks with x and y components each of which is discretized into 11 bins. *Top:* Button loss curves for all 12 buttons. *Middle:* Left joystick loss curves for X and Y components of the joystick. This joystick controls character movement. *Bottom:* Right joystick loss curves for X and Y components of the joystick. This joystick controls the camera.