Research Statement

Cameron Allen
camallen@berkeley.edu
https://camallen.net

Motivation

My research goal is to steer artificial intelligence towards human flourishing. This requires studying both the computations that drive intelligent behavior and the alignment protocols that ensure AI systems do what we want. In both areas, **my work addresses a fundamental challenge that has existed since the earliest days of AI: identifying an appropriate conceptual *frame.*** A frame is what *enables* decision making; it is a mental model, a representation, that defines the problem to be solved. Almost all AI algorithms and alignment techniques implicitly assume a fixed conceptual frame as a starting point. **I argue that AI systems need to construct *their own* frames, and modify them as necessary, to ensure that their behavior remains aligned with human preferences.**

Problem framing comes so naturally to humans that we typically do not realize we are doing it at all. We automatically convert our long-term preferences about possible futures (e.g. a career as a professor) into concrete, actionable goals (publish papers, mentor students, apply to faculty jobs). We identify high-level actions that seem appropriate for achieving those goals (write research statement, recruit letter writers). We transform our rich perceptual observations (raw sensory data) into mental state representations that summarize the most important information for selecting actions (current status of application materials). We map all of this (goals, actions, and states) onto previously learned behaviors and/or reasoning strategies (technical communication, planning, prioritization of subtasks). And then we move on to another preference (being fluent in French) and repeat with a different frame.

Finding an appropriate frame is essential for long-term human flourishing and is also most of what makes AI problem solving difficult. The traditional strategy of specifying a fixed frame by hand will only get us so far, because, as humans, we often don't know what to specify. While some of our preferences are easy to articulate, other preferences (e.g. ideal 2035 global AI policy) are nearly impossible to assess from our present context and will likely change over time. As AI systems rapidly become more capable, they are increasingly accomplishing our mis-specified objectives faster than we have time to correct them [1], [2]. We are already seeing that any fixed frame we specify may eventually be wrong, and a perfect specification of human flourishing remains quite far off. To safely support us in our efforts, AI systems must learn how to adapt. **This calls for a research program in *frame-aware AI,*** to develop techniques for AI systems to inspect, modify, and select between conceptual frames, while interactively learning (and helping us discover) our uncertain, changing human preferences.

Overview

My research program studies conceptual frames in AI systems to develop techniques that improve both AI and human decision making. One of my recent papers made **the first substantial progress in decades at determining when an AI system needs memory**—i.e. automatically detecting when a conceptual frame is *missing information* [3]. Another showed the **first evidence that a chess-playing neural network had *learned* to simulate future moves** before acting—effectively discovering a new conceptual frame *for itself* that enhanced its capabilities [4]. A third line of work introduced methods for learning human-interpretable state representations [5] and human-like high-level actions that precisely manipulate them [6]. The latter directly **led to an explainable tutoring system for teaching Rubik's cube to high school students** [7]. I am also **building systems that interact with real people**, including: a **code generation assistant** that maintains uncertainty about user goals rather than assuming it knows them exactly; a project with Mozilla and Northeastern to **mitigate security risks from AI-driven web browsing** in Firefox; and collaborations with social

and cognitive scientists at Berkeley and NYU adapting frame-aware AI methods to better understand creativity and memory in humans.

My work has resulted in: **9 conference papers** (plus 4 more under review) and **12 workshop papers** —at venues like NeurIPS, ICML, AAAI, IJCAI, AISTATS, ICAPS, and RLC; **2 oral/spotlight talks**; and **14 invited talks**. These qualifications demonstrate my ability to lead a strong research program focusing on AI frame inspection, modification, selection, and alignment.

Frame Inspection

Neural networks already construct internal conceptual frames—often discovering representations researchers did not explicitly program. In NeurIPS 2024 work [4], my colleagues and I showed that chess-playing networks internally simulate board states 3+ moves into the future before selecting moves, revealing that **planning representations emerge automatically** from pattern-matching training. This matters because there has been significant debate about the degree to which frontier language models (which use the *same neural architecture* as the chess network we studied) internally implement principled optimization algorithms like planning. Our work demonstrates that when the internal frame is allowed to drift, free from human feedback, **the AI may develop complex algorithmic behaviors we didn't account for**—a capability that could one day pose dangerous risks to humans [8].

Beyond studying neural representations "in the wild", I also train simplified models under carefully controlled experimental conditions and study the representations that implicitly emerge. I recently co-authored a workshop paper [9] that showed neural networks encode relational data as geometric structure. This work **enables an entirely new class of interpretability tools** for inspecting how neural networks represent *functions*, as opposed to just scalars, while also providing a strong theoretical explanation of how these representations are encoded in the neural network. I also supervised a workshop paper [10] looking at the extent to which recurrent neural networks trained via reinforcement learning internally represent their state uncertainty and perform Bayesian belief updates based on their observations. This is a useful tool, both for validating algorithms that learn what to remember, and for **revealing model uncertainty or overconfidence and taking appropriate safety precautions**.

Frame Modification

Observation Abstraction

AI agents operating in complex environments need compressed, interpretable representations. My dissertation work [11] developed methods that automatically convert the agent's rich sensory observations to concise, abstract state representations that **provably preserve problem-solving ability.** Rather than allowing full freedom for the agent's representation to drift, I identified a set of sufficient theoretical conditions and encoded them into the neural network objective function so it would learn representations compatible with the agent's learning algorithm. This demonstrates a method for enabling AI-driven frame modification while ensuring the resulting representation retains the properties that human designers want.

Subsequently, I supervised two extensions. The first uses the AI agent's actions to separate its representations into **distinct, human-interpretable state variables** [12]. The second leverages predefined state variables (and skills for manipulating them) to automatically build discrete, symbolic representations, **enabling the use of highly efficient classical planners** that would otherwise be incompatible with low-level continuous control [13]. These projects demonstrate how AI systems can

progressively construct new, abstract conceptual frames, entirely on their own, while constraining them to be increasingly human-interpretable.

Action Abstraction

Alongside observation abstraction, which shifts the AI agent's frame on the input side, I also study action abstraction, which alters the agent's outputs. In an IJCAI 2021 paper, I developed a method [6] that constructs a library of skills whose effects are intentionally focused on a small subset of state variables. Limiting side-effects in this way makes the skills **highly interpretable**, and it also leads to **orders of magnitude faster planning**, since focused skills better align with the assumptions of the available search heuristics. My work directly enabled **an explainable tutoring system** led by researchers at USC and Cisco **for teaching Rubik's cube to high school students** [7]. More recently, I advised two interns on a follow-up project [14] that generalized these ideas from deterministic planning to stochastic reinforcement learning, thereby providing a method to learn exactly the kinds of abstract actions required in the preceding section for constructing factored and symbolic representations. These projects suggest that action and observation abstraction naturally complement each other, and that **human-interpretable abstractions are beneficial for human and AI learning alike.**

World Models

Establishing a suitable frame defines an interface through which the agent can observe and model the world. The latter introduces a *modeling* problem, requiring a new choice of frame: the class of world model to consider. Whatever the agent chooses, for a complex enough world, the model will eventually need to change. One of my current projects [15] is on simplifying complex world models while **provably preserving plan validity and optimality** with respect to a particular planning goal. Another [16], applies discrete diffusion over programmatic world models to **efficiently propose edits when models make incorrect predictions**. These projects achieve complementary objectives: constructing *equivalent* world models that are more *efficient*, and constructing *modified* world models that are more *accurate*. The former ensures that newly proposed conceptual frames are actually useful for within-frame optimization. The latter allows for quick recovery when models diverge from reality.

Frame Selection

My NeurIPS 2024 paper [3] made **the most tangible progress in nearly 30 years at automatically detecting when AI agents need memory.** In our approach, an agent maintains two different value function models, with differing assumptions about the (in)completeness of its sensor observations, and checks whether the models agree to provably detect missing information. Moreover, we showed how to *learn* a memory—i.e. an abstraction over history—to expand the agent's representational frame with more context until the two models ultimately agree. This project demonstrates how an AI system can make meta-cognitive decisions about *which* frame is most appropriate for decision making, and how to automatically adjust its frame to match the learning algorithm.

This approach is quite general, and the same method could be used ensure human decision-makers have sufficient historical information to make accurate predictions (e.g. forecasting medical outcomes from patient history). I subsequently supervised two follow-up projects: one formalizing this notion of viewing memory as history abstraction [17]; and another [18] generalizing the notion of value discrepancies to arbitrary functions of observations besides just rewards, thereby decoupling the problem of frame selection from that of value optimization. I recently began collaborating with cognitive scientists at NYU to investigate whether similar value function discrepancies play a role in human behavior.

Frame Alignment

I am now supervising several projects on integrating frame-aware AI into systems that interact with real people. One project [19] reframes **coding assistants** as cooperative *assistance games* where models maintain goal uncertainty, which enables the assistant to ask clarifying questions rather than carelessly generating code by assuming sufficient knowledge of user intent. Another project [20] is a collaboration with Mozilla and Northeastern to bring **AI-driven web browsing to Firefox** while monitoring agent actions for privacy-violating behaviors and intervening before sensitive data is transmitted. A third [21] is a collaboration with the Berkeley Institute of Personality and Social Research to understand human creativity through the lens of conceptual frames. We argue that "creative" actions are those that lead to more accurate frames, and the value of such actions cannot be predicted in advance within the old frame.

Future Directions

The traditional AI paradigm optimizes for a fixed set of known human preferences and is already showing signs of strain. Even recent work in large language models, which aims to learn the human preferences from feedback during pretraining, still deploys a model with a fixed objective that is *known to be wrong*. My work in assistance games builds on earlier research to define a better frame, one in which AI agents optimize for *unknown* human preferences. Assistance games constitute a substantial improvement over fixed-preference approaches in that the deployed AI system **maintains uncertainty about what the human wants**, and learns how to act over time. I intend for frame-aware AI to go beyond either of these approaches, by allowing the frame itself to be learned over time.

The first step in my proposed plan would be to investigate more general-purpose methods for detecting when a conceptual frame is misaligned. I am already advising on a project [22] that uses probabilistic dependency graphs to explicitly model inconsistent beliefs about the world, extending Bayes Networks, which constrain beliefs to be consistent by assumption. This flexibility allows an AI system to **automatically detect modeling inconsistencies** and provides a richer language for model edits and updates to its beliefs, particularly when new evidence doesn't fit with predictions.

The next step is to **connect frame modification directly to human preferences**. This is challenging for several reasons. First, as humans, we do not know our "true" preferences over all possible futures, because we have not lived through them and do not have sufficient information for comparison. Second, the preferences we *think* we have are an approximation to our true preferences, and they change over time as we update them to account for new information. Third, we struggle to even *articulate* our preferences accurately, despite massive collective effort (such as in language model pretraining). However, my research portfolio offers plenty of reason for hope.

My plan is to explicitly represent the human's "true" preferences using a sufficiently rich set of preference factors (similar to the ones from my work learning factored representations). These "true" preferences would depend on an unknown subset of these features, and the human would try to estimate which features are important over time through interactions with the world. Meanwhile, the AI system would have its own estimate about both the preference features and the human's preference model, in essence solving a double assistance problem. On the preference level, it would solve an assistance game, and on the feature level, it would be solving a frame consistency problem (similar to my value discrepancy work, or my current project in probabilistic dependency graphs).

I have already shown how AI systems can learn to construct their own frames, as well as how to ensure those frames remain human-interpretable and aligned with downstream learning algorithms. This research program would combine these ideas together, **to build frame-aware systems that interact with real people and improve collaborative human-AI decision making at scale.**

References

[1] J. Becker, N. Rush, E. Barnes, and D. Rein, "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity," *arXiv preprint arXiv:2507.09089*, 2025.

[2] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S.R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S.R. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang, and E. Perez, "Towards Understanding Sycophancy in Language Models," *arXiv preprint arXiv:2310.13548*, 2023.

[3] C. Allen, A. Kirtland, R.Y. Tao, S. Lobel, D. Scott, N. Petrocelli, O. Gottesman, R. Parr, M.L. Littman, and G. Konidaris, "Mitigating Partial Observability in Sequential Decision Processes via the Lambda Discrepancy," in *Advances in Neural Information Processing Systems*, 2024.

[4] E. Jenner, S. Kapur, G. Vasil, C. Allen, S. Emmons, and S. Russell, "Evidence of Learned Look-Ahead in a Chess-Playing Neural Network," in *Advances in Neural Information Processing Systems*, 2024.

[5] R. Rodriguez-Sanchez and G. Konidaris, "Learning Abstract World Models for Value-preserving Planning with Options," *Reinforcement Learning Journal*, 2024.

[6] C. Allen, M. Katz, T. Klinger, G. Konidaris, M. Riemer, and G. Tesauro, "Efficient Black-Box Planning Using Macro-Actions with Focused Effects," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 2021, pp. 4024–4031.

[7] K. Lakkaraju, V. Khandelwal, B. Srivastava, F. Agostinelli, H. Tang, P. Singh, D. Wu, M. Irvin, and A. Kundu, "Trust and ethical considerations in a multi-modal, explainable AI-driven chatbot tutoring system: The case of collaboratively solving Rubik's Cube," in *ICML Workshop on Neural Conversational AI (TEACH)*, 2023.

[8] E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant, "Risks from Learned Optimization in Advanced Machine Learning Systems," 2019.

[9] J. Yocum, C. Allen, B. Olshausen, and S. Russell, "Neural Manifold Geometry Encodes Feature Fields," in *NeurIPS Workshop on Symmetry and Geometry in Neural Representations*, 2025.

[10] J. Liévano-Karim, P. Koepernik, G. Konidaris, and C. Allen, "Echo of Bayes: Learned Memory Functions Can Recover Belief States," in *NeurIPS Workshop on Unifying Representations in Neural Models*, 2025.

[11] C. Allen, N. Parikh, O. Gottesman, and G. Konidaris, "Learning Markov State Abstractions for Deep Reinforcement Learning," in *Advances in Neural Information Processing Systems*, 2021, pp. 8229–8241.

[12] R. Rodriguez-Sanchez, C. Allen, and G. Konidaris, "Disentangling Independently Controllable Factors in Reinforcement Learning," in *New York Reinforcement Learning Workshop*, 2025.

[13] A. Ahmetoglu, S. James, C. Allen, S. Lobel, D. Abel, and G. Konidaris, "Skill-Driven Neurosymbolic State Abstractions," in *Advances in Neural Information Processing Systems*, 2025.

[14] J. C. Carr, Q. Sun, and C. Allen, "Focused Skill Discovery: Using Per-Factor Empowerment to Control State Variables," *Reinforcement Learning Journal*, vol. 6, 2025.

[15] *In prep.*

[16] *In prep.*

[17] A. Kirtland, A. Ivanov, C. Allen, M. L. Littman, and G. Konidaris, "Memory as State Abstraction over Trajectories," in *6th Multidisciplinary Conference on Reinforcement Learning and Decision Making*, 2025.

[18] P. Koepernik, R. Y. Tao, R. Parr, G. Konidaris, and C. Allen, "General Value Discrepancies Mitigate Partial Observability in Reinforcement Learning," in *RLC Finding the Frame Workshop*, 2025.

[19] *In prep.*

[20] *In prep.*

[21] *In prep.*

[22] *In prep.*