Research Statement

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

Motivation

We are in the midst of a global artificial intelligence revolution that most people did not ask for and will not benefit from. Modern AI promises to speed us up, automate our drudgery, and make us more productive. Not only does it do the opposite [1], but it even cheers us on [2] as we steadily outsource more and more of our critical thinking. AI is supposed to make us more creative, but instead we are drowning in slop [3]. We face catastrophic risks on multiple fronts [4]: mass unemployment; extreme power concentration; out-of-control hyper-capitalism; erosion of trust; automated warfare; long-term personal and societal disempowerment. In short, **our society is hurtling towards systemic collapse.**

The problem with AI is tunnel-vision at every level. Our AI systems cannot see outside the fixed optimization-based problem frames that researchers define for them. Researchers cannot envision an AI paradigm other than optimizing a fixed objective. The objectives we specify rely on contextual assumptions that we do not state explicitly (or often even realize). We test AI systems on a narrow set of benchmark "evals" and equate good performance with complete readiness to deploy. We deploy AI within a wide range of dynamic contexts that we treat as equivalent and static. We then measure only the first-order outcomes and ignore any second- or third-order effects due to external feedback loops.

Steering away from disaster and towards **human flourishing will require humans and AI systems to reason meta-cognitively about our respective conceptual frames**—the mental models through which problems are understood and solved. AI must not only operate *within* a given frame, but also recognize the frame, make it explicit, propose alternatives, navigate between frames as context demands, and help humans do the same. This would constitute a major departure from the contemporary model of AI, which takes the frame for granted, and a much-needed reintegration of some of the earliest ideas in the field. Success has the potential to bring about unprecedented levels of self-awareness, creativity, and problem-solving ability in both humans and AI.

Overview

My research program studies conceptual frames in AI systems to develop techniques that improve both AI and human decision making. One line of work 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* [5]. I am now collaborating with cognitive scientists at NYU to look for similar memory formation mechanisms in humans. Another project 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 [6]. Another line of work introduced methods for learning human-interpretable state representations [7] and human-like skills that precisely manipulate them [8]. The latter directly **led to an explainable tutoring system for teaching Rubik's cube to high school students** [9].

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**. 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 working with Hortus AI to develop tools that **reveal implicit assumptions embedded in municipal AI systems** prior to deployment in communities.

Prior and current work

Frame inspection

Neural networks already construct internal conceptual frames—often discovering representations researchers did not explicitly program. In NeurIPS 2024 work [6], 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 models that learn to optimize without our knowledge could pose novel risks [10], and 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.

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 [11] 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 [12] 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 shifting

Observation abstraction

AI agents operating in complex environments need compressed, interpretable representations. My dissertation work [13] developed methods that automatically construct abstract state representations while **provably preserving problem-solving ability.** Subsequently, I supervised two extensions. The first uses the AI agent's actions to separate its representations into **distinct, human-interpretable state variables** [14]. The second leverages pre-defined 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 [15]. Together, these projects show how an AI system can progressively construct new, abstract reference frames, entirely on its own, by imposing increasing amounts of structure on its internal representations.

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 [8] 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. More recently, I advised two interns on a follow-up project [16] that generalizes 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. This suggests that action and observation abstraction could be interleaved or combined to construct increasingly abstract reference frame hierarchies.

World models

When world models make incorrect or inconsistent predictions, agents must quickly revise them. Another line of my research studies how to automatically shift between different conceptual models. One project [17] is on simplifying complex world models while **provably preserving soundness and optimality** with respect to a particular planning goal. Another [18], 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 [5] made **the most tangible progress in nearly 30 years at automatically detecting when AI agents need memory.** Agents maintain two different value function models, with differing assumptions about the (in)completeness of sensor observations, and check whether the models agree to provably detect missing information. The same method could also be used ensure human decision-makers have sufficient historical information to make accurate predictions (e.g. forecasting medical outcomes from patient history). 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. I subsequently supervised two follow-up projects: one formalizing this notion of viewing memory as history abstraction [19]; and another [20] 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 am also advising on a few other frame selection projects. One project on skill generalization [21] maintains multiple modeling hypotheses about how **previously learned skills can be efficiently reused**, then attempts to execute the skills in new contexts to reveal the correct model. Another project [22] uses probabilistic dependency graphs to provide a framework for explicitly modeling inconsistent beliefs about the world, rather than constraining 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.

Real-world applications

I am now integrating frame-aware AI into systems that interact with real people. One project 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 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. I am also working with cognitive scientists at NYU to **explore similarities between human and AI representations**, starting with looking for evidence of value function discrepancies during memory formation.

Future directions

I have been fortunate to make substantial progress on this research program already, but there are many interesting directions in which I hope to take it from here.

Education

AI tutoring systems currently optimize for quick answers, creating student dependency. Frame-aware AI can instead help to shift towards **building deeper understanding over time,** such that students

eventually come to rely *less* on AI. One concrete approach would be to train language models to predict student problem-solving behavior, then use interpretability tools to reveal misunderstandings. By identifying what conceptual frames emerge, we could **link students' mental models with pedagogical theories of misconceptions**, or compare against instructors' frames.

Such a system could be combined with an AI tutor that uses the identified reference frames to decide on interventions. Rather than responding with correct answers, the tutor could pose questions designed to help students identify and modify their frames. Initial interventions could be rule-based or trained via reinforcement learning, to respectively provide interpretable baselines or enable data-driven discovery of effective strategies. This parallels my focused macro-actions work [8], where interpretability of learned skills **directly enabled collaborative AI tutoring systems** [9] for teaching Rubik's cube solving, demonstrating how clear, localized effects help students recognize patterns and understand subgoal structure.

Civic AI

Municipal AI systems embed unstated assumptions about community priorities and often fail to account for social and ecological feedback loops. Tools that extract these assumptions enable stakeholders to evaluate during procurement whether proposed systems align with the needs of the community. I am collaborating with civic AI researchers at Hortus AI to develop methods for **improving government AI procurement** using interpretability methods and other frame inspection techniques. I plan to develop tools that extract and surface the implicit assumptions embedded in municipal AI systems—revealing which community needs are prioritized or what risk factors drive decisions.

My work on probabilistic dependency graphs provides a formal framework for representing conflicting stakeholder beliefs and measuring their inconsistency. However, extending PDG methods to *reconcile* conflicting beliefs by revising the graph remains open research. One approach: PDG-based AI systems detect inconsistent beliefs, then work with humans to resolve them. This advances frame-aware AI by **treating stakeholder perspectives as competing conceptual frames** that systems must represent explicitly and navigate between, **rather than collapsing into a single 'optimal' objective**—particularly important when certain groups have fewer resources to voice their perspectives. This builds on my Markov abstraction work [13], where comparing stakeholder-level and system-level models could identify when algorithmic abstractions collapse important community differences.

Conclusion

Understanding how people and AI systems can develop awareness of their own conceptual frames —**moving from being trapped in frames to recognizing, choosing, and transcending them**—has applications far beyond education and civic engagement. **My long-term interest is in AI that supports this meta-cognitive capacity more broadly.** The problem of steering towards human flourishing will continue to be a relevant research direction for as long as humans continue making decisions.

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] M. Read, "Drowning in Slop," *Intelligencer*, Sept. 2024, [Online]. Available: https://nymag.com/intelligencer/article/ai-generated-content-internet-online-slop-spam.html

[4] R. Uuk, C.I. Gutierrez, D. Guppy, L. Lauwaert, A. Kasirzadeh, L. Velasco, P. Slattery, and C. Prunkl, "A Taxonomy of Systemic Risks from General-Purpose AI," *arXiv preprint arXiv:2412.07780*, 2024.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] V. M. J. S. Evan Hubinger Chris van Merwijk and S. Garrabrant, "Risks from Learned Optimization in Advanced Machine Learning Systems," 2019.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] *In prep.*

[18] *In prep.*

[19] 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.

[20] 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.

[21] A. De Mello Koch, A. Bagaria, B. Huo, Z. Zhou, C. Allen, and G. Konidaris, "Learning Transferable Sub-Goals by Hypothesizing Generalizing Features," in *AAAI Workshop on Generalization in Planning*, 2025.

[22] *In prep.*