Scaling a Rational Approach to Cooperative Intelligence

Presidential Young Professorship (PYP) Grant Proposal

Tan Zhi Xuan

Assistant Professor, NUS Department of Computer Science

Contents

1 Executive Summary **1**

2 Aims / Objectives **1**
- AI agents that reliably assist individual users . . . 2
- AI agents that cooperate in teams, institutions, and societies . . . 2
- Infrastructure for human and AI cooperation . . . 2

3 Background & Significance **2**

4 Research Design & Methods **3**
- AI agents that reliably assist individual users . . . 3
  - Assistive Planning under Uncertainty . . . 3
  - Grounded Cooperative Dialogue . . . 3
  - Web Assistant / Web Agent Planning Language . . . 3
  - CoPlayers / Smart Cooperative NPCs . . . 4
- AI agents that cooperate in teams, institutions, and societies . . . 4
  - Multi-Resolution Multi-Agent Bayesian Modeling . . . 4
  - Norm Learning, Reasoning and Institutional Modeling . . . 4
- Infrastructure for human and AI cooperation . . . 4
  - Computational Models of Rational Deliberation and Negotiation . . . 4
  - AI-augmented Argumentation and Group Deliberation . . . 5

5 Milestones & Deliverables **5**

Scaling a Rational Approach to Cooperative Intelligence

Presidential Young Professorship (PYP) Grant Proposal

Tan Zhi Xuan

Assistant Professor, NUS Department of Computer Science

1 Executive Summary

Cooperation is a core aspect of human-like intelligence, and essential for AI agents that aim to assist human users or integrate with human teams and institutions. However, current approaches to building cooperative AI agents face significant challenges: Systems based on large language models (LLMs) struggle with reliable planning[1] and theory-of-mind reasoning,[2] while approaches grounded in rational probabilistic inference and decision-making have historically been limited by computational intractability[3] and the lack of accessible engineering frameworks. However, my research has demonstrated that it's possible to overcome these limitations, scaling a rational, model-based approach to cooperative intelligence that is built upon efficient algorithms and platforms for probabilistic programming [AISTATS'23] and model-based planning [SM Thesis].

This proposal outlines a research program that will scale a rational approach to cooperative intelligence even further, advancing this approach across three interconnected levels: **(i)** AI agents that reliably assist individual users in complex environments like the open Web; **(ii)** AI agents that can understand and operate with multi-agent teams, institutional structure, and social norms; **(iii)** Socio-technical infrastructure that promotes human and AI cooperation through frameworks for rational deliberation and alignment with shared normative principles. These advances will enable applications from AI web agents and AI co-players in video games to norm-adaptive autonomous vehicles and AI-augmented group deliberation.

The proposed research builds upon my past work, which has already achieved significant breakthroughs in the computational efficiency and practical applicability of model-based cooperative AI. These include Sequential Inverse Plan Search (SIPS), which provides a 100-1000x speed-up over previous baselines for Bayesian goal inference [NeurIPS'20]; Cooperative Language-Guided Inverse Plan Search (CLIPS), which enables pragmatic instruction following with appropriate uncertainty [AAMAS'24a]; and Norm-Augmented Markov games for efficient learning of cooperative norms [AAMAS'24b]. These innovations demonstrate the feasibility of addressing core theoretical challenges in human-AI cooperation, and developing practical solutions that achieve order-of-magnitude improvements over previous baselines.

Executing this program will require advancing core capabilities for human-like cooperation alongside improvements in engineering platforms, conceptual foundations, and real-world applications. My expertise in rational AI engineering — combined with interdisciplinary knowledge across cognitive science, philosophy, and AI alignment — uniquely positions me to tackle these challenges.

2 Aims / Objectives

This research program advances a framework for model-based cooperative intelligence across three interconnected scales: individual assistance, larger-scale cooperation, and societal infrastructure. Together, these objectives form a full-stack approach to building cooperative AI systems.

Objective 1: AI agents that reliably assist individual users. This objective focuses on developing AI agents capable of reliable goal inference and assistance in real-world applications. Current AI assistants often fail to accurately understand user intentions or reliably achieve user goals, especially when user instructions are ambiguous, or when tasks require long-horizon reasoning. To overcome these limitations, I aim to develop probabilistic programming and model-based planning frameworks for AI copilots, smart NPCs, and web agents that can robustly infer, understand, and achieve user goals. Unlike current approaches that rely heavily on unreliable and inefficient LLMs , this will enable AI assistants that infer posterior distributions over user goals from user actions and instructions, then plan safely and reliably to achieve those goals by reasoning over world models, all while using orders of magnitude less data and computation.

Objective 2: AI agents that cooperate in teams, institutions, and societies. This objective addresses the challenge of developing AI agents that can function effectively within larger social structures. While individual AI assistance is valuable, many contexts require coordination within teams of agents, and adherence to institutional roles and norms. My research will develop agents that understand and integrate with team dynamics, organizational hierarchies, and societal norms via multi-resolution multi-agent modeling. By modeling and learning about the social world at multiple levels of resolution, these agents will be able to cooperate appropriately in contexts such as traffic systems, human-robot teams, and digital workplaces, while maintaining appropriate levels of deference, initiative, and coordination.

Objective 3: Infrastructure for human and AI cooperation. This objective aims to create socio-technical infrastructure that enables effective human-AI cooperation at scale. Rather than focusing solely on agent capabilities, this work recognizes the need for systems that facilitate interaction, deliberation, and negotiation in human-AI ecosystems, so as to promote mutually beneficial outcomes and avoid AI-enabled conflict. To achieve this, I aim to advance the fundamental science of human deliberation and normative reasoning, grounding this work in rational models of human normativity[4] and computational frameworks for argumentation and negotiation.[5,6] This in turn will support the development of algorithms that aid human deliberation, and mechanisms that facilitate beneficial equilibria in the future AI economy.

3 Background & Significance

Cooperative intelligence represents an open frontier in AI research.[7] The ability of AI systems to reliably assist humans and cooperate with agents across multiple scales underpins a wide variety of AI applications[8] — from AI copilots and virtual assistants to autonomous vehicles and collaborative robots. Despite this centrality, current approaches to building cooperative AI systems face significant limitations that impede their reliability, efficiency, and safety at scale.

The dominant paradigm in AI assistance today relies heavily on Large Language Models (LLMs), which have demonstrated impressive capabilities in natural language understanding and open-ended reasoning.[9] Despite this, LLM-based systems exhibit critical shortcomings in reliability and safety, particularly in complex scenarios which require long-horizon planning[1,10,11] or sophisticated theory-of-mind reasoning.[2,12] A promising alternative with a long pedigree is ***rational AI***:[13–15] An approach to designing rational autonomous systems grounded in explicit probabilistic modeling and decision-making under uncertainty, including models of humans and their goals.[16,17] However, rational approaches to AI have historically been constrained by computational intractability.[3] Before my research, such approaches were often too slow for real-time interaction in complex environments, and lacked user-friendly engineering frameworks, preventing broader adoption despite their theoretical soundness.

During my PhD, I made significant progress in addressing these scaling challenges. Through innovations like Sequential Inverse Plan Search (SIPS), which achieves a 100-1000x speed-up in Bayesian goal inference over previous baselines [NeurIPS'20], Cooperative Language-Guided Inverse Plan Search (CLIPS), which enables uncertainty-aware instruction following by combining the linguistic abilities of LLMs with sound planning and inference algorithms [AAMAS'24a], and Norm-Augmented Markov Games, a Bayesian norm learning framework with orders of magnitude greater sample efficiency than model-free RL [AAMAS'24b], I have shown that model-based approaches to cooperative AI can be both theoretically grounded and practically efficient. My development of the PDDL.jl automated planning library [SM Thesis] and extensions to the Gen probabilistic programming system [AISTATS'23] has also significantly reduced the engineering barriers to developing planning algorithms and sequential Monte Carlo methods. These advances provide a foundation for the research program outlined in this proposal.

The significance of this research extends beyond technical innovation to address pressing societal concerns about AI alignment and safety. By advancing the science and engineering of cooperative intelligence across different scales — from individual interaction to institutional coordination — this research will contribute to building AI systems that not only assist humans more effectively, but that are aligned with normative principles and constraints that our society agrees upon [PhilStudies'24].

4 Research Design & Methods

My research program is structured around four categories of work: **(1) Cooperative Capabilities** that enable AI systems to cooperate with humans and other agents across multiple scales; **(2) Engineering Platforms** that increase the accessibility of model-based approaches to developing and deploying cooperative AI; **(3) Theoretical and Conceptual Foundations** that provide rigorous understanding of human-like cooperative intelligence; and **(4) Applications** that demonstrate and validate these advances in practical contexts. By advancing work across all four categories, this research program will deliver both fundamental insights and practical systems that embody those insights, achieving each of the objectives identified in Section 2.

Objective 1: AI agents that reliably assist individual users

To enable the development of model-based AI assistants, work on this objective will be structured around research projects advancing core cooperative abilities for real-time assistance in grounded environments, alongside platforms and applications for assistive AI:

- **Assistive Planning under Uncertainty** [*Capabilities*]: Assistive agents have to take actions under uncertainty about the goals of human users, infer subtasks to achieve, and gather information about both the environment and user goals. To address this, I aim to develop algorithms for real-time cooperative planning in belief space, building upon prior work in pragmatic goal assistance [AAMAS'24a], belief-space planning,[18] and Bayesian task delegation.[19] This work will also leverage advances in tractable probabilistic programming[20,21] to enable compact belief approximations and efficient belief updates, thereby overcoming tractability issues in belief-space planning.
- **Grounded Cooperative Dialogue** [*Capabilities, Platforms*]: Unlike chat-based AI assistants, assistive AI agents need to communicate reliably with users about actions, objects, and goals that are *grounded* in an external environment (e.g. web apps, video games, or the physical world). In this line of work, I aim to build upon the Bayesian instruction following paradigm in CLIPS [AAMAS'24a], developing AI agents that can ask targeted questions, describe the goals they are pursuing, and provide information that users may not be aware of. To achieve this, I plan to develop extensions of probabilistic programming systems for constrained generation from LLMs[22,23] with my collaborators at MIT, enabling uncertainty-aware semantic parsing and goal-oriented language production at scale.

- **WebAssistant / Web Agent Planning Language** [*Platforms, Applications*]: Current LLM-based web-browsing agents suffer from reliability[10] and safety issues[11] due to their lack of coherent world models[24] and sound planning abilities.[1] As an alternative approach to building AI web agents that is reliable-by-design, this project will develop the Web Agent Planning Language (WAPL), a domain-specific language for modeling web environments at the level of resolution needed for planning and inverse planning, taking inspiration from my prior work on PDDL.jl [SM Thesis]. WAPL will in turn enable the development of WebAssistant, a model-based, language-instructable AI web agent.
- **CoPlayers / Smart Cooperative NPCs** [*Platforms, Applications*]: While video-games have long made use of AI planning techniques like Goal-Oriented Action Planning,[25] cooperative planning between non-player characters (NPCs) and human players has been limited due to the absence of fast inverse planning technology. This project will bring SIPS-based inverse planning [NeurIPS'20] and grounded cooperative dialogue to video games via a game engine library for AI coplayers, while addressing the scaling challenges involved in real-time cooperative gameplay.

Objective 2: AI agents that cooperate in teams, institutions, and societies

Work on this objective will center around efficient multi-agent modeling, and the advancement of cooperative capabilities required for agents to interact with structured social environments:

- **Multi-Resolution Multi-Agent Bayesian Modeling** [*Capabilities, Platforms*]: Recent progress in multi-agent interaction has been driven by highly efficient simulators for multi-agent RL, which generate large amounts of synthetic data to train model-free policies. However, model-based cooperative agents require a different approach: Approximate but robust and efficient models of multi-agent behavior suitable for online Bayesian inference and planning. This project seeks to develop a multi-resolution multi-agent Bayesian modeling framework (MMBayes) that meets these needs, drawing upon ideas from SIPS [NeurIPS'20] and coarse-to-fine methods in probabilistic programming.[26] MMBayes will enable the development of autonomous systems (e.g. AVs, robots) that can rapidly and fluidly coordinate with large numbers of agents, modeling users and teammates at a high resolution while maintaining coarser-grained representations of less important agents (e.g. agents in crowds).
- **Norm Learning, Reasoning and Institutional Modeling** [*Capabilities, Foundations*]: For AI agents to integrate well within teams, institutions, and society at large, they have to understand the normative and institutional structure of their social environments. However, most prior work on social cognition in AI has focused on modeling the mental states of agents,[2,8] not the norms they collectively practice, or their social and institutional roles. To address this gap, this project will develop new theories, models, and algorithms for norm learning, reasoning and institutional modeling, building upon my work on Bayesian norm learning in Markov games [AAMAS'24b], game theoretic accounts of institutions[27] and norms,[28] and the cognitive science of institutional representations.[29] These models will be integrated into the MMBayes framework to enable richer multi-agent reasoning.

Objective 3: Infrastructure for human and AI cooperation

To foster fair and mutually-beneficial arrangements between humans and between AI agents, work on this objective will advance the fundamental theory and science of how humans negotiate and form agreements with each other, and apply this theory to design AI systems that embody and facilitate such interactions:

- **Computational Models of Rational Deliberation and Negotiation** [*Foundations, Capabilities*]: For AI systems to be aligned with human normative principles, they have to understand how and why people come to endorse certain principles, and whether those principles are reasonably agreed upon by relevant stakeholders [PhilStudies'24]. To design AI systems capable of replicating and augmenting

such processes, I aim to develop computational models that capture the rich structure of human deliberation and normative reasoning. This project will build upon work by myself and collaborators on resource-rational moral cognition,[4,30] integrating such work with ideas from computational argumentation frameworks,[5,6] game theory,[31] and social choice.[32]

- **AI-augmented Argumentation and Group Deliberation** [*Capabilities, Applications*]: As a demonstration and testbed for the computational frameworks described above, I aim to develop AI systems that facilitates group deliberation via argument analysis, summarization, and consensus generation. A specific use case is scientific peer review, where reviewers present arguments for and against the merit of particular paper, and a meta-reviewer facilitates discussion before producing a final judgment. Developing a system to produce such meta-reviews (AutoMeta) will provide a practically useful tool, while enabling the computational study of real-world human deliberation.

5 Milestones & Deliverables

Through this research program, I aim to deliver several software prototypes and research platforms that demonstrate the practical and theoretical benefits of model-based cooperative intelligence:

- **WebAssistant**: A model-based AI web agent built on CLIPS and the Web Agent Planning Language.
- **CoPlayer**: A library for language-instructable smart NPCs in cooperative video games.
- **MMBayes**: An multi-resolution Bayesian modeling framework for multi-agent societies.
- **AutoMeta**: A meta-reviewing AI system to facilitate peer review and group deliberation.

Below is a proposed research timeline, including intermediate milestones such as papers and tech demos:

Year	Milestones
Y1	• Recruitment of initial team of RAs / PhD students / post-docs with relevant skill sets • Initial design and prototype of Web Agent Planning Language (WAPL) • Establish partnership with game developers to create CoPlayer prototype • Paper on domain-general open-ended Bayesian goal inference • Paper on foundations and theory of norm-augmented Markov games
Y1.5	• Paper on grounded cooperative dialogue in language-augmented assistance games • Paper on efficient belief-space assistive planning
Y2	• Open-source WebAssistant + WAPL implementation as a web browser extension + paper • Initial theory and experiments on rational argumentation modeling
Y2.5	• Initial release and game demo of CoPlayer technology + paper • Paper on multi-resolution multi-agent Bayesian modeling / MMBayes prototype • Establish collaborations for applications of MMBayes
Y3	• Mature release of CoPlayer library • Version 1.0 of MMBayes, with collaborator-sourced use cases (e.g. self-driving, etc.) • Extending WebAssistant capabilities via automated WAPL modeling, etc.
Y3.5	• AutoMeta open-source prototype on real-world deliberation + peer review data • Theory paper on reason-based deliberation, negotiation, and norm-governed cooperation
Y4+	• Further extensions of existing software and modeling frameworks • Further work towards a unified theory of human-like cooperative intelligence

Key Papers & References

[NeurIPS'20] T. Zhi-Xuan, J. Mann, T. Silver, J. Tenenbaum, and V. Mansinghka, **"Online Bayesian Goal Inference for Boundedly Rational Planning Agents,"** *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[CogSci'21] A. Alanqary*, G. Z. Lin*, J. Le*, T. Zhi-Xuan*, V. K. Mansinghka, and J. B. Tenenbaum, **"Modeling the Mistakes of Boundedly Rational Agents within a Bayesian Theory of Mind,"** in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 43, 2021.

[CogSci'24a] T. Zhi-Xuan, G. Kang, V. Mansinghka, and J. Tenenbaum, **"Infinite Ends from Finite Samples: Open-Ended Goal Inference as Top-Down Bayesian Filtering of Bottom-up Proposals,"** in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 46, 2024.

[AAMAS'24a] T. Zhi-Xuan*, L. Ying*, V. Mansinghka, and J. B. Tenenbaum, **"Pragmatic Instruction Following and Goal Assistance via Cooperative Language-Guided Inverse Planning,"** in *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, 2024.

[SM Thesis] T. Zhi-Xuan, **PDDL. jl: An Extensible Interpreter and Compiler Interface for Fast and Flexible AI Planning**. SM Thesis, Massachusetts Institute of Technology, 2022.

[AISTATS'23] A. K. Lew, G. Matheos, T. Zhi-Xuan, M. Ghavamizadeh, N. Gothoskar, S. Russell, and V. K. Mansinghka, **"SMCP3: Sequential Monte Carlo with probabilistic program proposals,"** in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2023.

[CogSci'24b] L. Ying*, T. Zhi-Xuan*, L. Wong, V. Mansinghka, and J. Tenenbaum, **"Grounding Language about Belief in a Bayesian Theory-of-Mind,"** in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 46, 2024.

[TACL'25] L. Ying*, T. Zhi-Xuan*, L. Wong, V. Mansinghka, and J. Tenenbaum, **"Understanding Epistemic Language with a Language-augmented Bayesian Theory-of-Mind,"** in *Transactions of the Association for Computational Linguistics* 2025.

[AAMAS'24b] N. Oldenburg and T. Zhi-Xuan, **"Learning and Sustaining Shared Normative Systems via Bayesian Rule Induction in Markov Games,"** in *Proceedings of the 23rd Int'l Conference on Autonomous Agents and Multiagent Systems*, 2024.

[PhilStudies'24] T. Zhi-Xuan, M. Carroll, M. Franklin, and H. Ashton, **"Beyond Preferences in AI Alignment,"** *Philosophical Studies*, 2024.

Other References

[1] S. Kambhampati, K. Valmeekam, L. Guan, M. Verma, K. Stechly, S. Bhambri, L. P. Saldyt, and A. B. Murthy, "LLMs can't plan, but can help planning in LLM-modulo frameworks," in *Forty-First International Conference on Machine Learning*, 2024.

[2] H. Kim, M. Sclar, T. Zhi-Xuan, L. Ying, S. Levine, Y. Liu, J. B. Tenenbaum, and Y. Choi, "Hypothesis-driven theory-of-mind reasoning for large language models," *arXiv preprint arXiv:2502.11881*, 2025.

[3] D. Malik, M. Palaniappan, J. Fisac, D. Hadfield-Menell, S. Russell, and A. Dragan, "An efficient, generalized bellman update for cooperative inverse reinforcement learning," in *International Conference on Machine Learning*, pp. 3394–3402, PMLR, 2018.

[4] S. Levine, N. Chater, J. B. Tenenbaum, and F. Cushman, "Resource-rational contractualism: A triple theory of moral cognition," *Behavioral and Brain Sciences*, pp. 1–38, 2024.

[5] L. Amgoud and C. Cayrol, "A reasoning model based on the production of acceptable arguments," *Annals of Mathematics and Artificial Intelligence*, vol. 34, pp. 197–215, 2002.

[6] I. Rahwan, S. D. Ramchurn, N. R. Jennings, P. McBurney, S. Parsons, and L. Sonenberg, "Argumentation-based negotiation," *The Knowledge Engineering Review*, vol. 18, no. 4, pp. 343–375, 2003.

[7] A. Dafoe, Y. Bachrach, G. Hadfield, E. Horvitz, K. Larson, and T. Graepel, "Cooperative AI: Machines must learn to find common ground," *Nature*, vol. 593, no. 7857, pp. 33–36, 2021.

[8] K. M. Collins, I. Sucholutsky, U. Bhatt, K. Chandra, L. Wong, M. Lee, C. E. Zhang, T. Zhi-Xuan, M. Ho, V. Mansinghka, *et al.*, "Building machines that learn and think with people," *Nature human behaviour*, vol. 8, no. 10, pp. 1851–1863, 2024.

[9] A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, *et al.*, "OpenAI o1 system card," *arXiv preprint arXiv:2412.16720*, 2024.

[10] S. Zhou, F. F. Xu, H. Zhu, X. Zhou, R. Lo, A. Sridhar, X. Cheng, T. Ou, Y. Bisk, D. Fried, U. Alon, and G. Neubig, "WebArena: A realistic web environment for building autonomous agents," in *The Twelfth International Conference on Learning Representations*, 2024.

[11] I. Levy, B. Wiesel, S. Marreed, A. Oved, A. Yaeli, and S. Shlomov, "ST-WebAgentBench: A benchmark for evaluating safety and trustworthiness in web agents," *arXiv preprint arXiv:2410.06703*, 2024.

[12] H. Kim, M. Sclar, X. Zhou, R. Bras, G. Kim, Y. Choi, and M. Sap, "FANToM: A benchmark for stress-testing machine theory of mind in interactions," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14397–14413, 2023.

[13] S. J. Russell and P. Norvig, *Artificial Intelligence: A modern approach*. USA: Prentice-Hall, Inc., 1995.

[14] S. J. Russell, "Rationality and intelligence," *Artificial intelligence*, vol. 94, no. 1-2, pp. 57–77, 1997.

[15] S. Omohundro, "Rational artificial intelligence for the greater good," in *Singularity hypotheses: A scientific and philosophical assessment*, pp. 161–179, Springer, 2013.

[16] C. Baker, R. Saxe, and J. Tenenbaum, "Bayesian theory of mind: Modeling joint belief-desire attribution," in *Proceedings of the Annual Meeting of the Cognitive Science Society, 33 (33)*, 2011.

[17] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan, "Cooperative inverse reinforcement learning," in *Advances in neural information processing systems*, pp. 3909–3917, 2016.

[18] O. Macindoe, L. P. Kaelbling, and T. Lozano-Pérez, "Pomcop: Belief space planning for sidekicks in cooperative games.," in *AIIDE*, 2012.

[19] S. A. Wu, R. E. Wang, J. A. Evans, J. B. Tenenbaum, D. C. Parkes, and M. Kleiman-Weiner, "Too many cooks: Bayesian inference for coordinating multi-agent collaboration," *Topics in Cognitive Science*, vol. 13, no. 2, pp. 414–432, 2021.

[20] F. A. Saad, M. C. Rinard, and V. K. Mansinghka, "SPPL: Probabilistic programming with fast exact symbolic inference," in *Proceedings of the 42nd acm sigplan international conference on programming language design and implementation*, pp. 804–819, 2021.

[21] F. Zaiser, A. Murawski, and C.-H. L. Ong, "Exact Bayesian inference on discrete models via probability generating functions: A probabilistic programming approach," *Advances in Neural Information Processing Systems*, vol. 36, pp. 2427–2462, 2023.

[22] A. K. Lew, T. Zhi-Xuan, G. Grand, and V. K. Mansinghka, "Sequential Monte Carlo steering of large language models using probabilistic programs," *arXiv preprint arXiv:2306.03081*, 2023.

[23] J. Loula, B. LeBrun, L. Du, B. Lipkin, C. Pasti, G. Grand, T. Liu, Y. Emara, M. Freedman, J. Eisner, R. Cotterell, V. Mansinghka, A. K. Lew, T. Vieira, and T. J. O'Donnell, "Syntactic and Semantic Control of Large Language Models via Sequential Monte Carlo," in *The Thirteenth International Conference on Learning Representations*, 2025.

[24] K. Vafa, J. Chen, A. Rambachan, J. Kleinberg, and S. Mullainathan, "Evaluating the world model implicit in a generative model," *Advances in Neural Information Processing Systems*, vol. 37, pp. 26941–26975, 2024.

[25] J. Orkin, "Agent architecture considerations for real-time planning in games," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 1, pp. 105–110, 2005.

[26] M. Cusumano-Towner, B. Bichsel, T. Gehr, M. Vechev, and V. K. Mansinghka, "Incremental inference for probabilistic programs," in *Proceedings of the 39th ACM SIGPLAN Conference on Programming Language Design and Implementation*, pp. 571–585, 2018.

[27] G. K. Hadfield and B. R. Weingast, "What is law? A coordination model of the characteristics of legal order," *Journal of Legal Analysis*, vol. 4, no. 2, pp. 471–514, 2012.

[28] H. Gintis, "Social norms as choreography," *Politics, Philosophy & Economics*, vol. 9, no. 3, pp. 251–264, 2010.

[29] J. Jara-Ettinger and Y. Dunham, "The institutional stance," *Under review*, 2024.

[30] D. Trujillo, M. Zhang, T. Zhi-Xuan, J. B. Tenenbaum, and S. Levine, "Resource-rational virtual bargaining for moral judgment: Toward a probabilistic cognitive model," *Topics in Cognitive Science*, 2024.

[31] K. Binmore, *Natural Justice*. Oxford university press, 2005.

[32] S. Fish, P. Gölz, D. C. Parkes, A. D. Procaccia, G. Rusak, I. Shapira, and M. Wüthrich, "Generative social choice," *arXiv preprint arXiv:2309.01291*, 2023.