RESEARCH STATEMENT

Xuan (Tan Zhi Xuan)

How can we build AI systems that reliably assist humans with their goals, despite uncertainty about what those goals are? More broadly, how can we design machines to exhibit **cooperative intelligence**: machines that can align themselves with the objectives of individual users, coordinate with other agents to achieve shared goals, and comply with the norms and principles of society-at-large?

My research answers these questions through the framework of **Bayesian inverse planning**: By modeling humans as approximately rational agents that take actions or communicate instructions to achieve their goals, AI systems can infer distributions over people's goals from what they say or do, then act to help even under uncertainty. Drawing upon the tools of *probabilistic programming* and *model-based planning*, along with the insights of *computational cognitive science*, my research has shown (**1**) how Bayesian inverse planning algorithms can be engineered to run in (faster than) real-time while scaling to open-ended contexts with hundreds of possible goals. I have also shown (**2**) how these algorithms can integrate large language models (LLMs) as modeling subcomponents, enabling them to infer and disambiguate human intentions from ambiguous natural language. The (**3**) software platforms I have developed along the way have also allowed myself and my colleagues to study cooperative intelligence beyond goal inference, paving the way towards machines that can model not just our goals and plans, but also (**4**) our beliefs, values, and norms.

1 Real-Time Open-Ended Goal Inference via Bayesian Inverse Planning

People rapidly infer the goals of others by observing their actions over time. How can we design algorithms for goal inference that match this speed, while ensuring accuracy and calibration? During my PhD at MIT, I developed an algorithm class called **Sequential Inverse Plan Search (SIPS)**, which scales Bayesian inverse planning to run in (faster than) real-time (Figure 1). SIPS can process *10–75 actions per second* on a single CPU, while being more accurate and *100–1000 times faster than* inverse reinforcement learning (IRL) baselines [NeurIPS'20].

The key insight behind SIPS is that human planning can be modeled using *online* or *anytime* planning algorithms, which only plan ahead to a limited extent before committing to an action. This means that SIPS can tractably compute the likelihood of an agent's action given a goal, and hence rapidly update a posterior distribution over a set of goals. In contrast, classical plan recognition algorithms have to solve NP-hard planning problems when modeling what an agent will do, and standard IRL algorithms require costly RL inner loops or considerable offline training. SIPS also naturally accounts for *boundedly-rational* human planning: People might make mistakes and even fail to achieve their goals due to not planning ahead enough. By modeling this, SIPS can robustly infer others' goals even when their plans lead to failure, matching the flexible inferences of human observers in ways that standard goal inference algorithms cannot [CogSci'21].

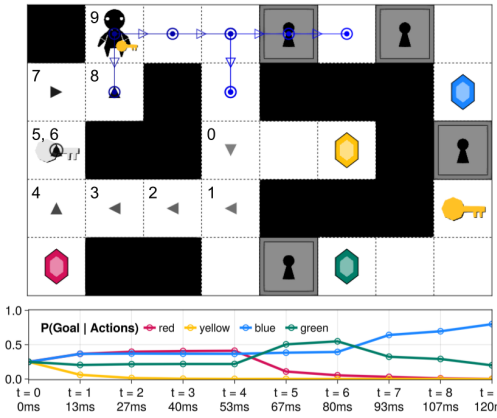

Figure 1: SIPS infers goals (gems) from actions (movement ▶, picking up items ◯) in real-time (13.3ms/step) by modeling agents that interleave depth-bounded anytime planning (blue search tree ▷⊙) with acting.

SIPS is not only fast, but also highly configurable—a benefit of its implementation in Gen,[1] a probabilistic programming framework that I co-maintain. Leveraging this flexibility, I have extended SIPS to perform *open-ended goal inference* in compositional goal spaces where there may be hundreds or more possible goals. Instead of exhaustively considering each goal, **open-ended SIPS** [CogSci'24a] makes use of *learned bottom-up proposals* that rapidly propose plausible goal hypotheses given the subgoals achieved by the agent so far. These bottom-up guesses are then filtered by importance reweighting and resampling, ensuring that the filtered hypotheses are good explanations of the observed actions. As a result, open-ended SIPS achieves runtimes *as low as 0.08s per action, and matches the accuracy of exact Bayesian inference over as many as 800 goals*. This addresses a key bottleneck in Bayesian approaches to cooperative AI, enabling us to scale uncertainty-aware AI assistance.

2 AI Assistants that Reliably Infer Intentions from Ambiguous Instructions

When working on shared tasks, people do not just signal their intentions through actions—they also communicate them in natural language. To handle such communication, I developed a language-based extension of SIPS called **Cooperative Language-Guided Inverse Plan Search (CLIPS)** [AAMAS'24a], which infers a human's goals and intentions from *both* their actions and instructions, then acts to assist that human. As a Bayesian algorithm, CLIPS is able to handle under-specified instructions, using observed actions as context to interpret ambiguous language, while providing safe assistance in cases of uncertainty. For example, a person might place three plates on the dining table, then say "Can you get the forks and knives?" Like humans, a CLIPS assistant can infer that the person's goal is to set the table for three, and hence plan to get three forks and three knives (Figure 2). If the person does not get any plates before making their request, then a CLIPS assistant will be appropriately uncertain about their goal, allowing it to ask clarifying questions.

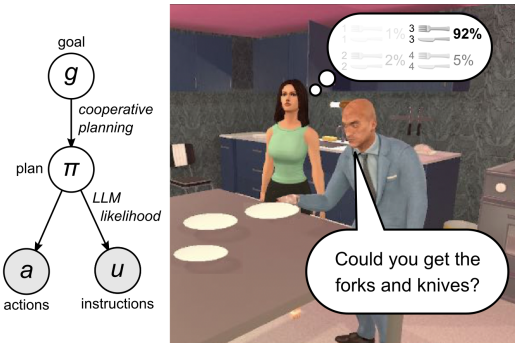

Figure 2: CLIPS performs multimodal goal inference from actions (a) and ambiguous instructions (u) by using LLMs as utterance likelihoods in a Bayesian agent model.

To achieve these inferences, CLIPS uses LLMs as sub-components of a probabilistic program, exploiting them as *utterance likelihoods* $P(u|\pi)$ over natural language inputs u given a symbolic plan π. Inverting this model, we can infer the human's intended plan π from an utterance u. CLIPS thus leverages the competence of LLMs at parsing a rich variety of natural language,[2] while avoiding their unreliability at cognitive tasks like planning[3,4] and theory of mind.[5,6] This allows CLIPS to be *2.5 times as accurate at goal inference than the multimodal language model GPT-4V*, despite CLIPS using a much smaller LLM (6.7B). The inferences and assistive plans produced by CLIPS were also highly similar to those of human raters (Pearson's r=0.93, vs. r=0.11 for GPT-4V). CLIPS thus paves the way towards language-based AI assistants that reliably infer and act according to our intentions.

3 Platform Engineering for Model-Based Planning and Programmable Inference

Modular software infrastructure is required to engineer systems like SIPS and CLIPS. Such infrastructure enables researchers to rapidly implement, debug, and iterate upon the models and algorithms involved. To address this need, I developed software platforms for model-based planning and programmable Bayesian inference, designing them for composability through well-defined APIs (Figure 3). This allows planning and inference algorithms to be combined with each other (as in SIPS), or with other AI technologies.

Efficient and domain-general planning algorithms are a key component of SIPS. To implement such algorithms, I created the **PDDL.jl interpreter and compiler** [SM Thesis] for planning tasks specified in the Planning Domain Definition Language (PDDL),[7] along with an ecosystem of planning algorithms (SymbolicPlanners.jl) and other tools. Unlike earlier planning software, PDDL.jl is designed around a *core interface* of operations used in model-based planning (e.g. querying whether a logical formula is true at a world state), and provides *multiple implementations* of that interface, including an interpreter, a compiler, and an abstract interpreter. This allows users to trade-off between debugging (via the interpreter) and speed (via the compiler), or to compute planning heuristics via abstract interpretation.[8] PDDL.jl also supports *extensible semantics*, allowing users to specify planning tasks which may involve numeric or array variables. Together, these features mean that PDDL.jl is both faster and more domain-general than most other automated planning systems, achieving run-times up to *36 times as fast* as the Pyperplan and ENHSP planners, and *within an order of magnitude* of the state-of-the-art FastDownward planner, while supporting *a wider range of planning domains* (classical, numeric, etc.) than all of these systems combined. The ease

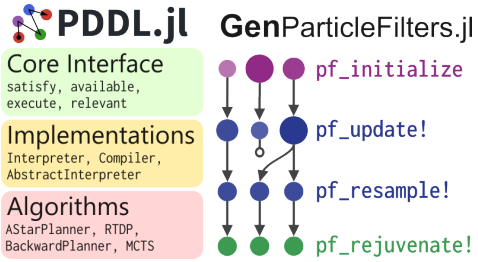

Figure 3: *PDDL.jl*: Efficient interfaces for AI planning. *GenParticleFilters.jl*: Modular building blocks for SMC.

of integration afforded by PDDL.jl has also led to growing adoption by the community (78 GitHub stars), and enabled innovative combinations of machine learning with planning heuristics by other researchers.[9]

Another key component of SIPS is its Sequential Monte Carlo (SMC) inference algorithm, which can be configured with custom resampling and rejuvenation strategies to effectively manage hypothesis diversity (as exploited by open-ended SIPS). To support this level of programmability, I created **GenParticleFilters.jl, a particle filtering and SMC inference library** for the Gen.jl probabilistic programming system. Like Gen itself, GenParticleFilters.jl is designed to support *programmable inference*,.[1,10] As such, it provides a wide variety of building blocks that developers can readily compose into full SMC samplers, including subroutines for particle initialization, updating, resampling, rejuvenation, and resizing. This makes GenParticleFilters.jl the most feature-complete probabilistic programming library for SMC, as exemplified by its support for the SMCP^3, an algorithm class that I co-developed which subsumes most existing frameworks for SMC [AISTATS'23]. By automating the low-level math required to ensure the soundness of SMC, GenParticleFilters.jl enables rapid prototyping of SMC algorithms that are tailored to specific model classes, enabling applications including SIPS, active learning of Gaussian Process kernel structure[11] and online synthesis of probabilistic programs.[12]

4 Beyond Goal Inference: Towards AI that Learns Our Beliefs, Values, and Norms

Human social cognition encompasses more than just goal inference; we are also adept at inferring others' beliefs, and learning the values and norms of the communities we are part of. With my mentees and collaborators, I have recently begun to show how the toolkits of probabilistic programming and model-based planning can be applied to build AI systems with these capabilities.

In the case of beliefs, we have extended SIPS to perform *belief-space inverse planning*, jointly inferring the goals and beliefs of an agent without exhaustive POMDP solving. By combining this algorithm with the formal semantics afforded by PDDL-based planning, and the language-to-code translation abilities of LLMs, we developed the a **language-augmented Bayesian theory-of-mind** model that is capable of interpreting and evaluating natural language statements about others' beliefs [CogSci'24b, TACL'TBD]. Our work breaks new ground by providing a principled semantics of belief sentences that is grounded in rational inferences about other agents' minds. Practically, our model demonstrates how we can build AI that reliably tracks our beliefs and how they may come apart from reality. This competency is crucial for algorithms that provide assistance or correction when we are ignorant or misinformed (e.g. intelligent tutors), and is currently lacking in even the largest LLMs.[6]

As for norms and values, a long-held interest of mine has been to reverse engineer the cognitive capacities involved in human moral learning and reasoning. To that end, my early work involved building robotic systems that incrementally learn rule-based ownership norms,[13] and studying Bayesian models of how humans disambiguate social norms from individual desires.[14] With my current expertise in planning and inverse planning, I have recently revisited these topics. Together with a Masters student that I supervised, I developed a new Bayesian framework for social norm learning in the context of multi-agent RL called **norm-augmented Markov games** [AAMAS'24b] (Figure 4). Implementations of this framework allow agents to passively learn complex social norms from *6–7 orders of magnitude less experience than model-free RL approaches*, and to emergently coordinate upon and stabilize shared conventions. I have also contributed to research on how people reason about social and moral norms, co-developing probabilistic models of *universalization reasoning*[15] and *virtual bargaining*[16] that explain when people decide it is permissible to break a certain rule. Finally, I led a position paper, **Beyond Preferences in AI Alignment** [PhilStudies'24], arguing for the importance of modeling human normative and evaluative reasoning when building AI systems that are aligned with human values, in contrast to dominant preference-oriented approaches. This paper has since garnered substantial interest, resulting in guest lectures and invited talks at Stanford University and the Simons Institute for the Theory of Computing.

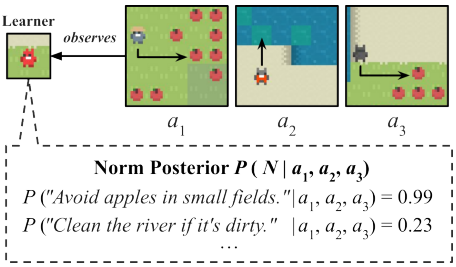

Figure 4: In Norm-Augmented Markov Games, agents infer cooperative norms from societal behavior.

5 Research Agenda: Scaling Cooperative Intelligence to Assistants, Teams, and Societies

Over the course of my PhD, I have laid the algorithmic and conceptual foundations for the design of cooperatively intelligent systems based on efficient Bayesian reasoning about other agents' minds. In the next phase of my career, I aim to build upon these foundations, scaling model-based cooperative intelligence towards **(i)** human-aligned AI assistants that operate safely and reliably in embodied and digital contexts by understanding the goals, beliefs, and intentions of their human principals, **(ii)** collaborative agents that automatically coordinate with humans and other agents by inferring shared goals and norms, **(iii)** algorithmically-enabled mechanisms that promote human cooperation, negotiation, and deliberation despite differences in our beliefs and values.

Human-Aligned AI Assistants for Physical and Virtual Worlds. AI assistants that take actions over long horizons in response to human instructions are growing rapidly in popularity, bolstered by the success of LLMs in processing open-ended natural language. However, due to the limitations of LLMs, such assistants remain too unreliable for widespread adoption, confabulating responses and taking unsafe actions without user approval. Algorithms like SIPS and CLIPS demonstrate that another path towards safe and reliable AI assistants is possible: By maintaining a coherent model of the physical or virtual world, as well as a model of the human user and their goals, CLIPS-style assistants can *ground the meaning of instructions in a well-defined semantics*, construct *plans with correctness guarantees*, and remain *appropriately uncertain* about user goals.

Going forward, I plan to *scale CLIPS to much richer spaces of goals, instructions, and environments*. To do so, I aim to leverage recent advances in probabilistic programming by my collaborators: Using SMC to enforce syntactic and semantic constraints on LLM outputs,[17,18] CLIPS will be able to parse natural language into complex temporal specifications. Using Bayesian 3D scene perception,[19,20] CLIPS will have access to environment models for planning and inverse planning. By integrating these technologies into a single stack, I believe it will be possible to create *digital agents and embodied robot assistants that are reliably human-aligned*.

Team-Forming and Norm-Abiding Agents for Human-AI Societies. As robots and AI agents become increasingly embedded in our societies, assisting individual humans will be just one of the capacities they require for cooperation and coordination. Building upon my work on team-based goal inference,[21] and Bayesian learning of social norms [AAMAS'24b], I plan to expand the science and engineering of agents that automatically coordinate with each other. Such coordination requires *inferring and jointly pursuing shared goals* (in cooperative settings), and *complying with norms and conventions that promote everyone's long-run interests* (in mixed-motive settings). These capabilities will be crucial to enable human-robot teams that collaborate fluidly in construction or rescue operations, autonomous vehicles that adapt or improve upon local driving conventions, and web-based AI agents that avoid over-taxing Internet infrastructure even while they pursue their users' goals.

Augmenting Human Cooperation by Enhancing Deliberation and Argumentation. People engage in deliberation, negotiation, and argumentation with each other in order to achieve shared or competing interests. Drawing upon our beliefs and values, we articulate reasons, assess empirical claims, and persuade others of the desirability of certain outcomes. Yet such interactions can break down, perhaps due to lack of factual agreement, failed understanding of each others' goals, or the inability to come up with mutually agreeable solutions. Leveraging my work on computational models of normative reasoning[15,16] and contractualist AI alignment [PhilStudies'24], I aim to develop *new models of human deliberation and argumentation* that are grounded in rational models of human cognition[22] and computational frameworks for negotiation and argumentation.[23] Informed by these scientific insights, I hope to design *algorithms that aid human deliberation* by combining formal models of negotiation and bargaining[24] with LLMs for open-ended text processing. Such algorithms would surface key areas of consensus and divergence, suggest ways to move beyond argumentative impasses, and propose solutions that fairly benefit all parties involved. If successful, this research program could transform how our societies get along, improving the quality and efficiency of our deliberation in contexts all the way from scientific peer review and contract negotiation to parliamentary and legislative debates.

Key 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'TBD] L. Ying*, T. Zhi-Xuan*, L. Wong, V. Mansinghka, and J. Tenenbaum, **"Understanding Epistemic Language with a Bayesian Theory-of-Mind,"** in *Transactions of the Association for Computational Linguistics* (Conditional Acceptance).

[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] M. F. Cusumano-Towner, F. A. Saad, A. K. Lew, and V. K. Mansinghka, "Gen: A general-purpose probabilistic programming system with programmable inference," in *Proceedings of the 40th acm sigplan conference on programming language design and implementation*, pp. 221–236, 2019.

[2] L. Wong, G. Grand, A. K. Lew, N. D. Goodman, V. K. Mansinghka, J. Andreas, and J. B. Tenenbaum, "From word models to world models: Translating from natural language to the probabilistic language of thought," *arXiv preprint arXiv:2306.12672*, 2023.

[3] I. Momennejad, H. Hasanbeig, F. Vieira Frujeri, H. Sharma, N. Jojic, H. Palangi, R. Ness, and J. Larson, "Evaluating cognitive maps and planning in large language models with CogEval," *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[4] 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.

[5] L. Ying, K. M. Collins, M. Wei, C. E. Zhang, T. Zhi-Xuan, A. Weller, J. B. Tenenbaum, and L. Wong, "The Neuro-symbolic Inverse Planning Engine (NIPE): Modeling probabilistic social inferences from linguistic inputs," *arXiv preprint arXiv:2306.14325*, 2023.

[6] 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.

[7] D. McDermott, M. Ghallab, A. Howe, C. Knoblock, A. Ram, M. Veloso, D. Weld, and D. Wilkins, "PDDL - the Planning Domain Definition Language," tech. rep., Yale Center for Computational Vision and Control, 1998.

[8] T. Zhi-Xuan, J. B. Tenenbaum, and V. K. Mansinghka, "Abstract interpretation for generalized heuristic search in model-based planning," in *ICML 2022 Workshop on Beyond Bayes: Paths Towards Universal Reasoning Systems*, 2022.

[9] L. Chrestien, S. Edelkamp, A. Komenda, and T. Pevny, "Optimize planning heuristics to rank, not to estimate cost-to-goal," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[10] V. K. Mansinghka, U. Schaechtle, S. Handa, A. Radul, Y. Chen, and M. Rinard, "Probabilistic programming with programmable inference," in *Proceedings of the 39th ACM SIGPLAN Conference on Programming Language Design and Implementation*, pp. 603–616, 2018.

[11] G. Z. Lin, "Bayesian active structure learning for Gaussian process probabilistic programs," Master's thesis, Massachusetts Institute of Technology, 2022.

[12] T. Mills, J. Tenenbaum, and S. Cheyette, "Human spatiotemporal pattern learning as probabilistic program synthesis," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[13] Z. X. Tan, J. Brawer, and B. Scassellati, "That's mine! learning ownership relations and norms for robots," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, (Accepted) 2019.

[14] Z.-X. Tan and D. C. Ong, "Bayesian inference of social norms as shared constraints on behavior," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 41, pp. 2919–2925, July 2019.

[15] J. Kwon, T. Zhi-Xuan, J. Tenenbaum, and S. Levine, "When it is not out of line to get out of line: The role of universalization and outcome-based reasoning in rule-breaking judgments," 2023.

[16] D. Trujillo Jiménez, M. Zhang, T. Zhi-Xuan, J. Tenenbaum, and S. Levine, "Resource-rational virtual bargaining for moral judgment: Towards a probabilistic cognitive model." Under review, 2024.

[17] 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.

[18] Anonymous, "Syntactic and semantic control of large language models via sequential monte carlo," in *Submitted to The Thirteenth International Conference on Learning Representations*, 2024. under review.

[19] N. Gothoskar, M. Cusumano-Towner, B. Zinberg, M. Ghavamizadeh, F. Pollok, A. Garrett, J. Tenenbaum, D. Gutfreund, and V. Mansinghka, "3dp3: 3D scene perception via probabilistic programming," *Advances in Neural Information Processing Systems*, vol. 34, pp. 9600–9612, 2021.

[20] N. Gothoskar, M. Ghavami, E. Li, A. Curtis, M. Noseworthy, K. Chung, B. Patton, W. T. Freeman, J. B. Tenenbaum, M. Klukas, *et al.*, "Bayes3D: fast learning and inference in structured generative models of 3D objects and scenes," *arXiv preprint arXiv:2312.08715*, 2023.

[21] L. Ying, T. Zhi-Xuan, V. Mansinghka, and J. B. Tenenbaum, "Inferring the goals of communicating agents from actions and instructions," in *Proceedings of the AAAI Symposium Series*, vol. 2, pp. 26–33, 2023.

[22] 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.

[23] 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.

[24] 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.