Research Statement

Vasanth Sarathy

Introduction

"Can you describe the shark?" This was a caption from Harry Bliss' New Yorker cartoon in which a lifeguard on a beach is yelling the question to a distressed swimmer in the ocean.[1] What would it take for an artificial intelligence (AI) agent to appreciate the humor in this cartoon? We humans effortlessly understand the absurdity involved here. Even though we do not see a shark, we assume that the swimmer is distressed because the shark is attacking them. In addition, we make a hidden assumption about social norms being violated - namely, that the lifeguard should be saving the swimmer rather than asking mundane questions.

Making and breaking these kinds of **assumptions** is crucial to human intelligence and creativity. They fill gaps in our knowledge and resolve ambiguities in an otherwise endless stream of perceptual information. When broken, they allow us to restructure our knowledge in new ways and discover previously unknown connections. Handling assumptions is required not only for understanding humor, but making sense of most everyday experiences and interactions. Future AI systems must also be able to ***reason with implicit assumptions*** if they are to work effectively with humans.

The problem, however, is that many of the necessary common sense assumptions are elusive and remain hidden in our collective minds. It is also unclear what kind of reasoning is needed to effectively use these assumptions while thinking.

Research agenda: Uncovering the computational nature of these hidden assumptions and how we make and break them are the central themes of my research program. In my research methodology, I aim to (1) identify AI domains that can benefit from reasoning with hidden knowledge, (2) formalize the types of knowledge and reasoning needed within these domains, (3) develop efficient algorithms for extracting the hidden assumptions, (4) develop cognitive architectures for reasoning with these assumptions, and (5) design studies to validate the effectiveness of new methods in practical settings. Developing effective approaches requires bringing together many disciplines covering areas of knowledge representation and reasoning, machine learning, natural language processing, uncertainty processing, human-robot interaction, as well as cognitive science, social psychology and ethics. Below, I discuss recent progress and outline future directions of research.

Common Sense Assumptions about Social Norms

One category of hidden common sense assumptions are those associated with human **social and moral norms**. Norms play an important role in society in maintaining order and ensuring cooperation and coordination. However, norms are challenging to computationalize as they are highly context-dependent, culture-dependent, and dynamic. Here, I will discuss my work with *modeling human norm acquisition and building AI agent architectures for norm learning and reasoning.*

[1] See back page for cartoon.

Social Norms Associated with Objects and Interactive Behaviors

Norms for object affordances: There is a tight connection between an object's physical appearance and the triggering of action possibilities in our minds. This allows us to quickly choose and manipulate an object correctly and effectively. When we see a knife, we also see possibilities for how and where we can grasp it. While visual and geometric features of objects might serve as useful cues for inferring basic actions like grasp, push, and pull, they do not tell us the complete story. For example, an antique knife behind a glass case in a museum does not allow for the same sort of grasping action. Other contextual aspects combined with our common sense social norms of the etiquette in a museum allow us to infer correctly. I introduced the formal notion of "**cognitive affordances**"[2] to capture these types of more complex normatively-charged action possibilities [10, 8]. In this work, I proposed a probabilistic logic-based formalism for reasoning about social norms associated with tools in a kitchen domain. I integrated this formalism into a cognitive robotic architecture that allowed for evaluating the performance on a PR2 robot [11, 9]. The robot was able to consider **implicit social norms** associated with handling sharp objects. It reasoned that when using a knife, it should grasp the handle, whereas when handing the knife over to someone, it should grasp the blade allowing for safe transfer.

Consent norms for robots: Certain normative assumptions extend beyond specific objects or domains and permeate all forms of interaction. One such example is the notion of "**implicit consent**," the idea that certain interactive behaviors demand prior implicit permissions. Consent (or lack thereof) can alter the normative valence of a situation, turning appropriate behaviors into inappropriate ones. In a recent paper, I turned to centuries-old legal precedent on consent in tort law and suggested how these established legal principles can guide robotics researchers while designing ethical behaviors and conducting usability studies [3].

Future work: I intend to continue capitalizing on my legal background and apply normative principles in specific domains where consent can influence robot behaviors. For example, I plan to identify the visual and linguistic consent cues suggestive of object ownership norms. Knowing when it is appropriate to borrow objects and tools is important for robots engaged in collaborative teamwork. I plan to design robotic architectures capable of detecting these cues and adjusting their behaviors to conform with these norms.

Generalized Norm Representation and Learning

Learning norms from language: Where do these social norm assumptions come from? One source is natural language instruction. I designed algorithms that allow a robot to learn, from natural language instruction received from single or multiple speakers, the social rules associated handling knives [6]. Not all norms are learned from language, however. Many are learned simply from observation. Observing "norms" is tricky as what one is observing is often not a norm per-se, but simply the performance (or non-performance) of an underlying behavior.

Modeling human norm learning: Along with my collaborators at Brown University, I explored the nature of human norms and their cognitive representations [2]. Crucially, we showed that these mental representations actually encode contextual aspects and inherently capture epistemic uncertainty [14]. We performed several human-subject studies and developed a preliminary computational model for human norm learning under uncertainty. The model was built on a logical framework and coupled with Dempster-Shafer theory (DST), a generalization of the Bayesian approach to learning under uncertainty. I extended this work to enable agents to represent their uncertainty over contexts as well [15].

Learning norms at scale: A limitation of DST is its computational complexity. I developed a set of generalized learning algorithms for reducing this complexity. I applied these algorithms to large-

[2] "Affordances" generally represent the relationship between an object and the set of actions an agent can perform on or with the object (e.g., a mug has a grasp affordance).

scale agent-based simulations during which an agent learned several norms from observing hundreds of other agents [5]. This work extended the current literature on multi-agent norm learning by showing how norms can be learned under more realistic conditions where the learning is influenced by incomplete, imperfect and unreliable data.

Future work: An advantage of DST is that it has the mathematical ingredients for "**open-world learning**," in which the space of possible behaviors (over which probability distributions are learned) is unknown at the start. An agent equipped with this capacity will be able to assimilate and establish the normativity of new behaviors on the fly while learning. I intend to design open-world agent architectures and algorithms that apply these powerful mathematical ingredients in DST and allow agents to learn from a state of complete ignorance not only of the normative status of behaviors, but of the knowledge of the behaviors themselves.

So, what do we do with these learned norms? In the above examples, the norms are immediately used for selecting the appropriate behaviors. Norms (and other hidden assumptions) can also be used for understanding natural language.

Nonmonotonic Reasoning for Language Understanding

Using hidden assumptions for resolving pronouns: Consider the discourse: "Smith entered the office of his boss. He was nervous."[3] We might assign the "he" pronoun in the second sentence to Smith assuming naturally that he is nervous because he is getting reprimanded, an interpretation that makes sense. Howevever, if our next sentence reads: "After all, he did not want to lose his best employee," we immediately revise our prior conclusion. This is an example of **nonmonotonic reasoning** and is pervasive in human language understanding. To handle such sentences, I introduced a new class of problems, which requires tracking state changes as language unfolds and reasoning about possible worlds within these intermediate states [13]. To resolve pronouns of this form, I proposed a multi-reasoner approach that employs Answer Set Programming, a declarative programming paradigm that permits nonmonotonic reasoning. I showed that current state of the art neural systems fail to correctly resolve even very seemingly simple discourses such as "Pick up block-A. Put it on block-B. Pick up block-C. Put it on block-B."

Using hidden assumptions for understanding speech acts: More recently, I extended the same architecture to resolving **indirect speech acts** (ISAs) [7]. These are utterances whose meaning might appear somewhat inconsistent with the surface form. For example, the question "Can you describe the shark?" might be requesting a shark description, and not asking the listener for a yes or no answer. Thus, the utterance has an indirect meaning of a request rather than an ask or a question. These are common politeness norms in human communication, and AI dialog systems must also be able to handle them correctly. However, much like the pronoun resolution problem mentioned earlier, ISAs also require nonmonotonic reasoning. I developed such a model that includes many domain-independent default rules and complex nonmonotonic reasoning patterns. I am currently developing a more challenging corpora for evaluating the model and interpreting ISAs in different contexts.

Future work: Like pronouns and ISAs, there are many linguistic phenomena (e.g., word sense and presuppositions) that are fraught with ambiguity and resolving which requires nonmonotonic reasoning with hidden knowledge. I plan to develop language datasets that can be used to jointly test and compare neural-based as well as logic-based approaches. I intend to use these datasets to evaluate the design of architectures for simulataneously resolving several forms of linguistic ambiguities, a capability offered by my multi-reasoner approach. I expect to use the reasoning and inference machinery that comes with logic-based approaches to extract hidden knowledge from text (e.g., online instructional datasets like WikiHow) as well as neural language models (e.g., BERT). I also plan to study domains, like legal reasoning, in which extracting the hidden knowledge is particularly

[3]This example is courtesy Grigris Antoniou.

challenging even for humans. In this domain, norm-based and nonmonotonic reasoning systems can help us evaluate the strength of legal arguments by extracting hidden assumptions.

Breaking Assumptions: From Common Sense to Creativity

Thus far, I have discussed the importance of common sense assumptions as gap fillers that allow us to make sense of situations. Sometimes, these common sense gap filling assumptions can misguide our perception of the world. Illusionistis and magicians frequently use this aspect of human cognition when designing tricks. Cartoonists, like Harry Bliss, also do so with drawings and text that seem incongruent with one another, but when combined suggest a better comic model. Human cognition is flexible and can break existing assumptions and adopt new ones in the face of new knowledge.

Formalizing creative problem solving: This occurs frequently in the real world, whether it is a mundane task of discovering a new affordance for an object (e.g., using a chair to hold open a door) or whether it is the historic task of a brilliant scientist proposing a paradigm shifting theory. In all these cases, we are (and can be) made to rethink our common sense assumptions and reframe them as needed. Undoubtedly, this is an important skill for AI systems as well, especially if they are to assist us with some of humanity's toughest challenges requiring major paradigm shifts. I recently proposed a class of problems, dubbed "**MacGyver Problems**" which are designed to seem initially unsolvable for an agent, but ones that an agent can conceivably solve by interacting with its environment [12]. I have also developed algorithms for solving certain classes of MacGyver problems [4]. MacGyver problems closely parallel insight problem solving, an area that has been studied by psychologists for over a century. I proposed a new cognitive model for creative problem solving that extends prior work on insight problem solving and is supported by neural evidence [1].

Future work: I plan to extend the theoretical framework for MacGyver problems by defining complexity classes and designing evaluation frameworks for testing new solution algorithms. The theoretical framework offers us the ability to systematize the process of generating novel stimuli for studying the neural basis of insight problem solving, an area ripe for collaboration with neuroscientists. Since this is a relatively new research area, I have promoted my research through public lectures, including giving a TEDx talk. For future funding, I co-authored a proposal to the NSF 2026 Idea Machine Competition. The competition aims to identify promising research directions and will set the NSF's research agenda for the next decade. My proposal for creative problem solving was selected as a finalist and is currently under review. If selected as a winner, it will open up an entirely new program that can fund numerous research projects in this area.

Conclusion and Funding

At the core of the process of making and breaking assumptions is the ability to continually assess and make sense of the world, resolving ambiguity to favor explanations that make the most sense and recognizing anomalies as possible avenues for creative exploration. Agents designed with the ability for sense-making are likely going to be more robust, resilient and resourceful. My work thus far has opened up many promising new research problems and provided me with a varied set of tools with which to explore them. I intend to secure funding for my exploration of these problems from governmental sources like the NSF, which has expressed interest as part of two programs – Robust Intelligence and Information Integration and Informatics – housed under the Department of Information and Intelligent Systems. Early in my career, I intend to secure an EAGER grant from the NSF to explore how hidden assumptions can be extracted from natural language texts. I intend to grow this work and secure a CAREER grant for pursuing more foundational work in human sense-making and creativity. In addition to sources like the NSF, I also intend to pursue additional sources of funding from the military (ONR, AFOSR, ARO) and industry partners.

References

[1] **Vasanth Sarathy**. Real World Problem-Solving. *Frontiers in Human Neuroscience*, 12, 2018.

[2] **Vasanth Sarathy**. Learning Context-Sensitive Norms under Uncertainty. In *Proceedings of the 2nd AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES-19)*, 2019.

[3] **Vasanth Sarathy**, Thomas Arnold, and Matthias Scheutz. When Exceptions Are the Norm: Exploring the Role of Consent in HRI. *ACM Transactions on Human-Robot Interaction (Formerly, Journal of Human-Robot Interaction)*, 8(3):14:1–14:21, July 2019.

[4] **Vasanth Sarathy**, Marlow Fawn, and Matthias Scheutz. On Solving Seemingly Impossible Problems. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, 2020.

[5] **Vasanth Sarathy**, Giordano Ferreira, Emily Sim, Matthias Scheutz, and Kamal Premaratne. Learning Context-Sensitive Norm Representations under Uncertainty. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, 2020.

[6] **Vasanth Sarathy**, Bradley Oosterveld, Evan Krause, and Matthias Scheutz. Learning Cognitive Affordances for Objects from Natural Language Instruction. In *Proceedings of the Sixth Annual Conference on Advances in Cognitive Systems*, 2018.

[7] **Vasanth Sarathy**, Antonio Roque, Alex Tsuetaki, and Matthias Scheutz. Interpreting Context-Sensitive Indirect Speech Acts using Non-Monotonic Reasoning. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, 2020.

[8] **Vasanth Sarathy** and Matthias Scheutz. Semantic Representation of Objects and Function. In *Proceedings of the 2015 IROS Workshop on Learning Object Affordances*, 2015.

[9] **Vasanth Sarathy** and Matthias Scheutz. Beyond Grasping - Perceiving Affordances Across Various Stages of Cognitive Development. In *Proceedings of the The Sixth Joint IEEE International Conference Developmental Learning and Epigenetic Robotics (ICDL)*, 2016.

[10] **Vasanth Sarathy** and Matthias Scheutz. Cognitive Affordance Representations in Uncertain Logic. In *Proceedings of the 15th International Conference on Principles of Knowledge Representation and Reasoning (KR)*, 2016.

[11] **Vasanth Sarathy** and Matthias Scheutz. A Logic-based Computational Framework for Inferring Cognitive Affordances. *IEEE Transactions on Cognitive and Developmental Systems*, 10(1):26–43, 2018.

[12] **Vasanth Sarathy** and Matthias Scheutz. MacGyver Problems: AI Challenges for Testing Resourcefulness and Creativity. *Advances in Cognitive Systems*, 6, 2018.

[13] **Vasanth Sarathy** and Matthias Scheutz. On Resolving Ambiguous Anaphoric Expressions in Imperative Discourse. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019.

[14] **Vasanth Sarathy**, Matthias Scheutz, Joseph Austerweil, Yoed Kenett, Mowafak Allaham, and Bertram Malle. Mental Representations and Computational Modeling of Context-Specific Human Norm Systems. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, 2017.

[15] **Vasanth Sarathy**, Matthias Scheutz, and Bertram Malle. Learning Behavioral Norms in Uncertain and Changing Contexts. In *Proceedings of the 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 2017.

"Can you describe the shark?"