HCRI RESEARCH SUMMARY

HCRI

HUMANITY CENTERED
ROBOTICS INITIATIVE
RHODE ISLAND

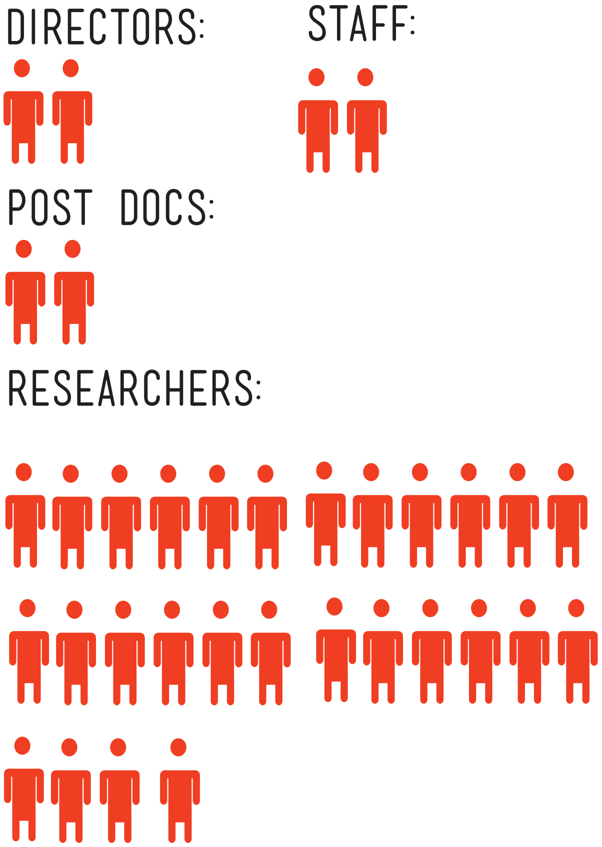
BROWN

HCRI AT A GLANCE

We are a group of Brown University faculty, students, and affiliates dedicated to robotics as a means to tackle the problems the world faces today. Beyond pursuing the goal of technological advancement, we want to ensure that these advancements are applicable and beneficial economically and socially. We are working across many disciplines to document the societal needs and applications of human-robot interaction research as well as the ethical, legal, and economic questions that will arise with its development. Our research ultimately aims to help create and understand robots that coexist harmoniously with humans.

The HCRI unites Brown University faculty and students across numerous departments and schools who are dedicated to robotics as an innovative and societally beneficial technology. Common commitments include (a) identifying societal needs that robots can help fulfill; (b) advancing science and technology of robots that fulfill these needs; and (c) studying and integrating into design the societal impact of robotic technologies, with a goal of averting labor replacement and privileged access to technology.

At the Brown University Humanity Centered Robotics Initiative we believe that the AI safety debate is skewed towards long-term threats, even though there are serious issues regarding AI safety facing us today. These include the use of opaque AI in criminal sentencing, personnel hiring, and access to loans, to name a few. Among other topics, we are working on ways of improving legibility in machine learning and AI, and improving communication with robots—physical embodiments of AI—that will be working in partnership with humans in the near future. We feel that solving the immediately pressing issues of AI safety will better position us to consider longer-term threats regarding human-machine interactions. Moreover, some of the solutions to the immediately pressing issues (e.g., human-guided machine learning, robots that are intelligible to humans and respond to correction) will provide solutions to the proximal threats to society imposed by AI.



RESEARCH AREAS

AI Explainability
Moral Norms
Human Interpretable Reinforcement Learning
Human-Robot Communication Through Augmented Reality
Decision Making in Self Driving Vehicles
Human Robot Teaming
UAV Energy Efficiency
Social Robotics for The Elderly
Social Robotics for Children with Type One Diabetes
Pick and Place Novel Object Detection
Motion Planning in Milliseconds
Infrared Fiducial Markers
Lightfield Arrays
VR Teleoperation

RESEARCH DETAIL 1: AI EXPLAINABILITY

Why do AI systems need to provide explanations?

Artificial Intelligent Systems (AIS) are rapidly entering human society, processing vast amounts of information and making classifications or recommendations that humans use for financial, employment, medical, military, and political decisions. The call for these systems to be transparent has recently become loud and clear (e.g., Wachter, Mittelstadt, and Floridi 2017). People want to understand how AIS operate so they anticipate, accept, and correct their decisions and determine whether the systems can be trusted. However, the decision making processes of highly complex AIS are often opaque to users, and sometimes even to creators. What is more, these systems have been found to engage in biased, unfair, and therefore untrustworthy decision making (Angwin, Larson, Mattu, & Kirchner, 2016), so it becomes imperative to understand where the biases come from.

To address this problem, scientists, developers, and researchers have intensified work on AIS transparency, but there are several different forms of transparency we need to distinguish. Some forms, such as traceability and verification, are particularly important for software and hardware engineers to test the internal structure and workings of an application. Other forms are particularly important for ordinary people and are typically called intelligibility or explainability—referring to the system's ability to make its operations intelligible, or better yet, to explain its decisions and actions in ordinary language. Current efforts are focusing on revealing the bases of classification decisions (e.g., Hendricks et al., 2016; Ribeiro, Singh, & Guestrin, 2016), but especially in human-machine interactions, it is the meaning of actions that requires explanation.

In the attempt to provide explanations of machine actions to humans, it is important to understand how people conceptualize AIS. When people interact with AIS, they inevitably construct mental models to understand and predict their actions (Epley, Waytz, and Cacioppo, 2007). However, people's mental models of AIS stem from their interactions with living beings. Thus, people easily run the risk of establishing incorrect or inadequate conceptions of these systems, which can result in people either under-trusting or over-trusting the system (Wortham, Theodorou, & Bryson, 2016). To prevent such miscalibrated trust, researchers and designers need to understand people's expectations and inferences about AIS and implement system behaviors that enable people to form appropriate mental models of the systems. Such models will allow people to better judge AIS capabilities, predict their intentions, understand their actions, and potentially correct their errors. This will increase not only acceptance of AIS but also facilitate calibrated trust in them. Thus, our research goals are to examine human mental models of AIS and use this understanding to design AIS features (e.g., describing and explaining their own actions) that genuinely allow people to understand the systems and adopt appropriate levels of trust (Ososky, Schuster, Phillips, & Jentsch, 2013).

How do we create AIS that explain? Three approaches

We take three approaches to the problem of creating explainable AIS, and the approaches will increasingly intertwine to offer a systematic solution to the need for AIS to be intelligible to humans.

The first approach seeks to describe and explain the sequential behaviors of a reinforcement learning (RL) system in a Markov Decision Process (MDP). We use linear temporal logic as the task-independent language that describes these sequential behaviors, and we use counterfactuals to identify the critical features that guide the learning algorithm's behavior, thereby helping explain its behavior.

The second approach helps a learning algorithm acquire explanations of its own behaviors by receiving models of such explanations from human teachers. The teachers draw the learner's attention to important aspects of its input, dependencies among the input, and offer examples of what good explanations are.

The third approach builds on our longstanding work on how humans explain behavior—specifically, what conceptual framework underlies such explanations and what vocabulary appropriately expresses these explanations. There is good reason to believe that people will use this same framework and vocabulary for AIS that they experience as "agents" (e.g., social robots), but currently there is no evidence to support this assumption, and we do not know under what conditions people might use alternate frameworks. We therefore provide the first systematic study of how people explain robots' and other artificial agents' behaviors. In addition, if people do use this conceptual framework to explain AIS behavior, they will expect AIS to explain their own behaviors in very similar ways. And because our previous work has laid out in detail the elements of this framework, we can begin to implement it in AIS and thus shape their explanations in ways that are maximally intelligible to humans.

Approach 1: Describing and explaining MDP behaviors

We propose an action-explanation approach made up of the following components:

A trained reinforcement learning (RL) system will be used to generate a variety of trajectories on a set of test problems. These trajectories will be fed to a learning algorithm that seeks to express the sequential behavior in a descriptive language such as linear temporal logic (LTL).

Existing work[1] has shown that such descriptions can be extracted using an optimization approach based on genetic programming. We believe a deep learning approach inspired by neural Turing machines[2] can scale more effectively and leverage recent improvements in continuous optimization. Our past work[3] provides a relaxation of logic that could be the basis of such an approach. Previous explanation work in Partially Observable Markov Decision Processes and Abstract Markov Decision Processes[4] has shown the computational limitations of state estimation based systems and have led us towards event-based frameworks such as LTL.

Explanations would then be synthesized by the system carrying out "what if" experiments. If the learner had been following a modified version of the extracted description, would it have behaved differently? If so, the part that was modified is a good basis for an explanation of the decision the system made.

We plan to apply this explanation system to a modern RL system and evaluate the resulting output with human judges to see if the explanations are perceived as insightful and satisfying. We will also examine whether the explanations have predictive and generalizable power, allowing human observers to guess the system's behavior in other (related) situations or future system states.

Approach 2: Humans teaching explanations to machines

Today's machine-learning systems, while incredibly powerful, can also produce bewildering outputs. In some cases, like AlphaGo's "Move 37", the decision[5] confounded even experts because it came from a completely alien analysis of a hard problem.

[1] https://pdfs.semanticscholar.org/fdf1/a3c114123d79b209fe295fa30918bd4043e2.pdf
[2] https://arxiv.org/abs/1410.5401
[3] https://arxiv.org/abs/1704.04341
[4] http://h2r.cs.brown.edu/wp-content/uploads/2017/03/whitney17.pdf
[5] https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/

RESEARCH DETAIL 1: AI EXPLAINABILITY

In other cases, like the responses of powerful learning algorithms[6] to “adversarial inputs”, people are surprised when an exceptionally well-trained machine fails spectacularly. AIS that can produce not only responses but also explanations of its responses are key to separating these two situations, giving us greater insight into when machines are misbehaving and when they can teach us something new.

Existing approaches to explanation generation are grafted on top of standard learning algorithms and, as such, can require just as much engineering to create as the systems they seek to explain. We propose an approach that turns explanation on its head by exploring novel training procedures in which human experts act as teachers. However, the teachers are not simply providing labels (e.g., “positive instance” vs. “negative instance”) but additionally offer examples of good explanations. This form of teaching includes basic kinds of explanations: drawing the learner’s attention to important aspects of its input, describing the dependencies between key input features, and illuminating the relationships between simpler and more complex concepts that the learner is grappling with. We believe that such a training procedure will enable machines to learn more complex concepts with fewer examples and, as a result, will help democratize machine learning. Google- or Amazon-scale companies will no longer need to create massive datasets to train powerful algorithms—anyone can train a machine to solve the problems they need help with.

In addition, machines trained using human explanations are in a considerably better position to explain their own decisions in a way that is understandable for ordinary people. They will have the vocabulary of explanations provided by their teachers, and they will have explicit representations of the key relationships that define their concepts. Instead of grafting an explanation module to the outer shell of the learning algorithm, explanations would permeate the entire system.

Approach 3: Identifying the properties of successful explanations

In addition to allowing systems to learn from the expressed explanations of human teachers, we also study in depth what concepts and vocabulary constitute these human explanations and how they constrain what will count as intelligible and successful explanations given by AIS. We have extensive work on the conceptual framework that underlies people’s explanations of human behavior and the kind of vocabulary that appropriately expresses these explanations (Malle, 2004, 2011). We will now apply this work to the question of what constitutes explanations of AIS behavior, explanations given both by humans and by machines themselves.

The conceptual framework of human behavior explanation consists of the fundamental distinction between intentional and unintentional behavior (Malle & Knobe, 1997) and the distinct explanatory tools people use for each type of behaviors (Malle, 1999). The primary explanation mode for intentional behaviors is by way of reason explanations, which cite an agent’s reasons for deciding to act. Reasons refer to desires and beliefs inferred about the agent and therefore open a window to the agent’s mind at the time of action. As a result, when evaluating explanations of AIS behavior, people will similarly expect to have a window into the AIS “mind” and understand the reasons for its intentions and actions. It would not suffice for the AIS to state, “I do this in order to maximize my rewards” or “Because it is the optimal policy.” Instead, when, say, a healthcare robot

[6] https://arxiv.org/pdf/1412.6572.pdf

declines the care recipient's re-quest for an increase in pain medication, it would have to say, "I am not allowed to change your pain medication without your doctor's consent, and I have not yet been able to reach her." Or when a hotel guest enters her room, and finds a robot circling the bed, the robot might say, "I hope I am not disturbing you; my duty is to tidy up your room." Indeed, studies show that people prefer explanations given by AIS that use this language of reasons (Harbers, van den Bosch, & Meyer 2009). Obviously, these early studies did not use machines that actually implemented the conceptual framework of behavior explanation, but that must be the ultimate, though extremely challenging goal[7].

Currently, we know very little about the conditions under which people apply this framework of explanation to machine behavior, and we do not know the functional benefits or risks of applying this framework (e.g., expecting reasons from a simple algorithm may lead to disappointment when the system actually reveals its inner workings). Thus, detailed insights into the scope and limits of people's humanlike treatment of AIS are needed. We will examine how people explain AIS actions and what such explanations reveal about the cognitive and social underpinnings of people's relations with these systems. Such findings will then guide design of AIS that explain their own behaviors, using the appropriate concepts and vocabulary that humans expect. Major research questions include: (1) What is needed for AIS to distinguish the major distinction between intentional and unintentional behavior and to apply the appropriately distinct modes of explanation? (2) What kind of vocabulary will AIS need to use to deliver appropriate explanations? (3) How can this vocabulary be coordinated with the system's actual underlying algorithms?

Integration among approaches

The three approaches will converge and complement one another in several ways. For example, we will try to translate the LTL vocabulary from Approach 1 into the conceptual framework that people use to describe and explain human behavior. That is, we will try to give the LTL formulas that explain MDP behaviors an interpretation in the language of beliefs, desires and intentions, and because people are familiar with this language MDP-based agents will become intelligible. In addition, because we know the elements of the framework of ordinary behavior explanation, we know what human teachers in Approach 2 are doing when they provide explanations for the learning agent. The elements of this framework could therefore be provided to the agent from the start so that it expects the right kinds of explanations and maximally benefits from its human teachers.

CS deliverables

1. A new algorithm for supervised learning that uses simple human explanations (region selection, labels of subcomponents) to learn more efficiently.
2. A new neural-net-based algorithm for extracting linear temporal logic expressions from observing the behavior of a trained reinforcement learner.
3. An approach for taking a decision taken by a trained reinforcement learner and a linear temporal logic description of its overall behavior and producing a "what if" style of explanation---the decision taken is explained by a specific part of the logical formula.

Psychology Deliverables

1. Systematic experiments that reveal the conditions under which humans apply their natural conceptual framework of behavior explanation to AIS. Conditions may include appearance (e.g., machine-like

[7] Unintentional behaviors, by contrast, do not stem from intentions and belief-desire reasons; instead, they are explained by reference to a wide variety of causes, such as physical obstacles, behaviors, circumstances, and the like. Such cause explanations of unintentional behavior are conceptually no different from cause explanations of physical events such as rolling rocks. In none of these cases does anybody or anything form an intention or adopt a reason. Thus, explanations of unintentional behavior will be less challenging to implement in AIS.

vs. humanoid), initial behavior (e.g., routine task behaviors or creative solutions), claimed capacities (e.g., “this is an autonomous intelligent robot” vs. “this is a robot”), domain of application (e.g., senior case vs. security check), and role (e.g., human collaborator vs. solow worker).
2. Documentation of the natural vocabulary that humans use to explain AIS behavior.
3. An evaluation of whether the resulting explanations give human participants insight into the behavior of the system in other related situations.
4. Experiments to examine whether AIS that use this vocabulary to explain their own behavior are more predictable, better understood, and elicit calibrated levels of trust.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. ProPublica. Retrieved July 14, 2017, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

de Graaf, M., & Malle, B. F. (in press). How people explain action (and autonomous intelligent systems should too). In Proceedings of the AAAI Fall Symposium Series (Symposium on Artificial Intelligence for Human-Robot Interaction), November 2017.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. Psychological Review, 114, 864–886.

Harbers, M., van den Bosch, K., & Meyer, J.-J. C. (2009). A study into preferred explanations of virtual agent behavior. In Z. Ruttkay, M. Kipp, A. Nijholt, & H. H. Vilhjálmsson (Eds.), Intelligent virtual agents (pp. 132–145). Berlin, Heidelberg: Springer.

Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., & Darrell, T. (2016). Generating visual explanations. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), Computer Vision – ECCV 2016., Lecture Notes in Computer Science, (pp. 3–19). Cham: Springer International.

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. Personality and Social Psychology Review, 3, 23–48.

Malle, B. F. (2004). How the mind explains behavior: Folk explanations, meaning, and social interaction. Cambridge, MA: MIT Press.

Malle, B. F. (2011). Time to give up the dogmas of attribution: A new theory of behavior explanation. In M. P. Zanna & J. M. Olson (Eds.), Advances of Experimental Social Psychology (Vol. 44, pp. 297–352). San Diego, CA: Academic Press.

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. Journal of Experimental Social Psychology, 33, 101–121.

Ososky, S., Schuster, D., Phillips, E., & Jentsch, F. G. (2013, March). Building Appropriate Trust in Human-Robot Teams. In AAAI Spring Symposium: Trust and Autonomous Systems.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16 (pp. 1135–1144). New York: ACM Press.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics.” Science Robotics 2 (6): eaan6080.

Wortham, R. H., Theodorou, A., & Bryson, J. J. (2016). Robot transparency, trust and utility.” In AISB Workshop on Principles of Robotics, 2016, University of Sheffield.

RESEARCH DETAIL 2: MORAL NORMS

This project rests on the assumption that Artificial Intelligent Systems (AIS) can become safe and beneficial contributors to human communities if they—like all beneficial human contributors—are able to represent, learn, and follow the norms of the community. We label this set of abilities **norm competence**, and we bring together social, cognitive, and computer sciences to advance research both on how such norm competence appears in the human mind and how it might be implemented in artificial minds.

Norms are not merely a subset of an agent's goals but rather constrain the agent's pursuit of goals. Were it not for the norms governing a particular context, an individual (human or artificial) would pursue a variety of maximally rewarding actions that might not be rewarding for other community members. Norm-guided action, in a sense, maximizes a societal value function. We argue that AI and robotics, too, should work to maximize societal value, and they can do so if norm competence lies at the foundation of AIS behavior.

We propose an ambitious program of experimentational and computational work on three core elements of human and artificial norm systems: Representation and activation of norms; learning of norms; and action implementation of norms. In all cases, we use novel experimental paradigms to study human norm representations, develop computational models of such representations, and begin to implement these models in artificial agents that thereby acquire norm competence.

For orientation, we offer a working definition of norms (Malle, Scheutz, & Austerweil, 2017):

Definition: *A norm N is an instruction to (not) perform an action A in context C, N:= C → D(A), provided that a sufficient number of individuals in a community (a) indeed follow this instruction and (b) demand of each other to follow this instruction.*

By the above definition, norm representations are sensitive to the practices and demands of communities. The latter feature distinguishes norm-guided actions from many other actions, such as those guided by collective desires or preferences: Even when most people in the community perform a certain action (e.g., singing in the shower), if there is no community demand to do so, the action does not reflect a norm.

For humans to act on the vast number of social and moral norms that a community imposes on them, those norms must be cognitively represented in efficient but flexible ways (e.g., in case community demands or practices change). Our program of work examines these cognitive representations of norms through innovative cognitive research, computational modeling, and interaction research that assesses the impact of norm-competent AIS on human-machine collaborations.

Major Research Questions

1. Representation and activation of norms

A first goal of this proposal is to identify the key properties of norm representation: how norms are cognitively organized in the human mind and activated in the contexts in which they apply. Knowing about these properties will guide design of norm systems in artificial agents that have to match the functional properties of human norm systems. The questions we try to answer include:

- How are norms structurally organized? Do they form networks of varying association strength, similar to semantic networks?
- What enables norms to be activated in the "right" kinds of contexts? Are norms triggered by specific features of the physical environment, especially cultural artifacts that tell the individual how culture and community want them to act?

- What properties of norms allow people to resolve conflicts between norms, through trade-offs and justifications of exceptions?

2. Learning and updating norms

Cognitive norm representations must be acquired and continuously updated, so our second goal is to identify principles of norm learning that humans rely on and that machines will have to rely on as well. We will investigate how humans and AIS can acquire new norms, integrate them with already acquired norms, and update the entire norm system. In particular, we will examine four processes that aid norm learning: (1) observations of the environment (cultural traces such as artifacts and signs); (2) observation of community members' behavior; (3) observation of community members' (dis)approval of other members' behavior; and (4) learning from explicit instruction. We will explore the ways in which representational properties of human norms (e.g., rapid context-specific activation, hierarchy) constrain these learning processes and how different kinds of norms (e.g., prohibitions vs. prescriptions) interact with the suitability of different learning processes.

3. Norm competence in human-machine collaboration

The third goal is to examine how norm competence improves human-machine collaboration. We test the following hypotheses. First, human-machine collaboration involving AIS with norm competence will be more efficient, because norms offer mutual predictability through a limited repertoire of acceptable and likely actions. Second, humans will better understand the machine's behavior, because it would be familiar, often well-adjusted to people's own behavior, and explicable not only by individual goals but by the norms that human and AIS share. Third, as a result of this predictability and understanding, humans will be more likely to have calibrated trust in AIS—trust that the machine will be reliable when it abides by relevant norms and distrust when it does not. People will feel safe and comfortable interacting with robots that share their norms and values.

Years 1-2: Representing norms

Supported by a DARPA grant, we recently developed experimental methods to extract community norms from ordinary people's responses to everyday scenes (Malle et al., 2017; Kenett, Allaham, Austerweil, & Malle, 2016). We presented people with numerous contexts (e.g., photo of a library or board room) and asked them to generate, as quickly as possible, actions that one is "supposed to do" here (for prescriptions) or "forbidden to do here" (for prohibitions). We found that these responses show robust patterns of consensus and context (see Figure 1). In a second set of studies, employing a signal-detection paradigm, we demonstrated that when given naturally generated norm- guided actions along with a depiction of the relevant context, people are more accurate and faster in recognizing the norms prescribed for the particular context than norms forbidden in that context.

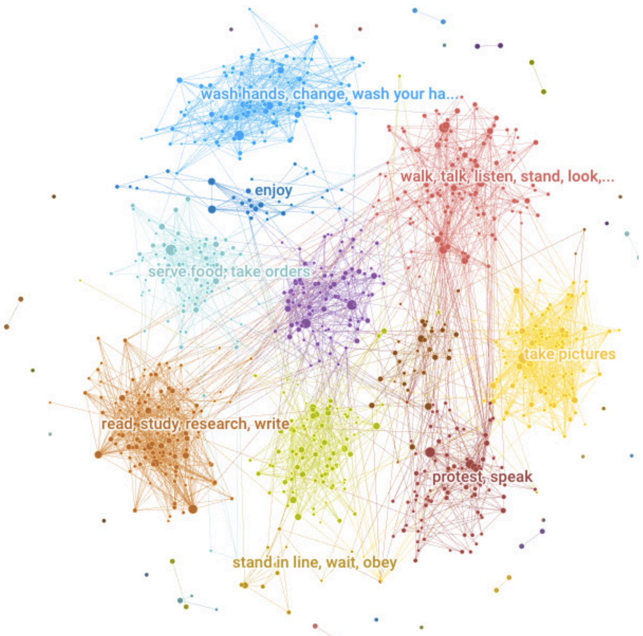


Figure 1:
A network representation of prescribed actions in eight distinct scenes (e.g., library, protest march, bathrooom), exhibiting high consensus and context specificity. Visualized using the Vibrant Data mappr tool.

RESEARCH DETAIL 2: MORAL NORMS

Building on this initial work, we plan to determine whether norms are represented and processed similarly to semantic networks and other standard knowledge structures or have unique properties that require distinct cognitive theories and computational models. One non-semantic theory is that norms are action programs directly activated by features afforded by the physical environment. These features will often be traces left by culture and community (e.g., others' behaviors, object arrangements, signs) that any agent, human or artificial, could learn. Such delegation to the environment would cut down on storage needs, but it raises the important question of how prohibitions are represented; for, in humans at least, activating the negation of an action program may be too dangerous (because it also activates the action program itself). So, prohibitions might have their own cognitive and computational structure (a possibility supported by our initial experiments), including inhibitory, not merely excitatory processes.

We will examine additional properties of norms that enable the human mind—and would enable an artificial agent—to activate appropriate norms in any given context. In particular, we will consider the community demand (the second part of our norm definition) that human agents experience and that artificial agents must expect and be sensitive to. Do human norm representations have a normative-force parameter that specifies how strongly the community demands obeyance of the given norm? And will formal implementations of norms in AIS be able to incorporate such a parameter?

We will combine classic experimental paradigms (using text, pictures, and videos) with newly available augmented and virtual worlds (using the Unity game engine operating on Hololens or Brown's virtual reality theater, the "Yurt"). In these worlds, we control the environment's norm-relevant features, including cultural artifacts, and measure participants' patterns of norm activation (by recording their verbal and nonverbal behaviors, reaction times, and eye movements) as well as responses to their own or others' norm violations (by measuring pupil dilation, skin conductive and other stress responses). From these activation patterns, we can infer properties of norm representations, such as context specificity and hierarchy, using recent developments in network science (e.g., Freeman, 1978; Newman, 2010).

With a cognitive and computational model of norm representation and processing in place, the next step is to build specific norm models for social domains in which robots and other AIS are likely to be deployed in the near future (e.g., settings of senior care, crowd security, or search and rescue). The models we construct will preserve the properties of human norm representation, such as hierarchy and rapid activation. We can then build a virtual robot with these norm models and assess its behavior in the virtual worlds we have previously validated with human participants, a focus of years 3 and 4 described below.

Psychology deliverables:

A series of experiments that document the properties of human norm representations, especially:
1. the scope and limits of context sensitivity,
2. the network characteristics of norm constellations,
3. the triggers of norm activation (e.g., cultural artifacts, gist inferences), and
4. the mechanisms of conflict resolution among norms activated in the same context.

CS deliverables:

1. Computational models of generic norm networks that reflect the properties of norms identified in the experimental work (e.g., high context-sensitivity, fast local activation, hierarchical resolution of norm conflicts).
2. Computational implementations of specific norm networks of select contexts (e.g., senior care, search and rescue), including testbed performance of the system suggesting appropriate actions and rejecting inappropriate actions under slight context variations.

RESEARCH DETAIL 2: MORAL NORMS

Years 3-4: Learning norms and human-machine collaboration

The norm-learning work will examine four fundamental mechanisms people use to acquire and update norms—mechanisms that artificial agents may have to use as well: (1) observations of the environment (cultural traces such as artifacts and signs); (2) observation of community members' behavior; (3) observation of community members' (dis)approval of other members' behavior; and (4) learning from explicit instruction.

Because the research literature on norm learning is sparse, we will develop novel experimental paradigms that control and vary features of the physical environment, the availability of community members in the scenes, and the availability of explicit instruction (e.g., commands, signs). We will capture people's default norm-learning tendencies in different environments, the efficacy and evidence strength afforded by different mechanisms, and the integration and reconciliation of data from different mechanisms. We will also explore the ways in which representational properties of human norms (e.g., rapid context-specific activation, hierarchy) constrain these learning processes and how different kinds of norms (e.g., prohibitions vs. prescriptions) interact with the suitability of different learning processes.

We will then apply our insights to develop algorithms with norm learning capacity. Our approach is variegated, leveraging the strength of machine learning algorithms from different traditions, such as reinforcement learning, logic-based, and hybrids as appropriate. To help AIS learn norms from observations of others' behavior, we will begin by applying the idea of inverse optimal control. That is, the learner can see what others do, but it cannot perceive what norms motivate that behavior. The framework of inverse reinforcement learning (IRL) is an appropriate starting point for inferring the norm-based motivations decision-makers follow. An IRL system

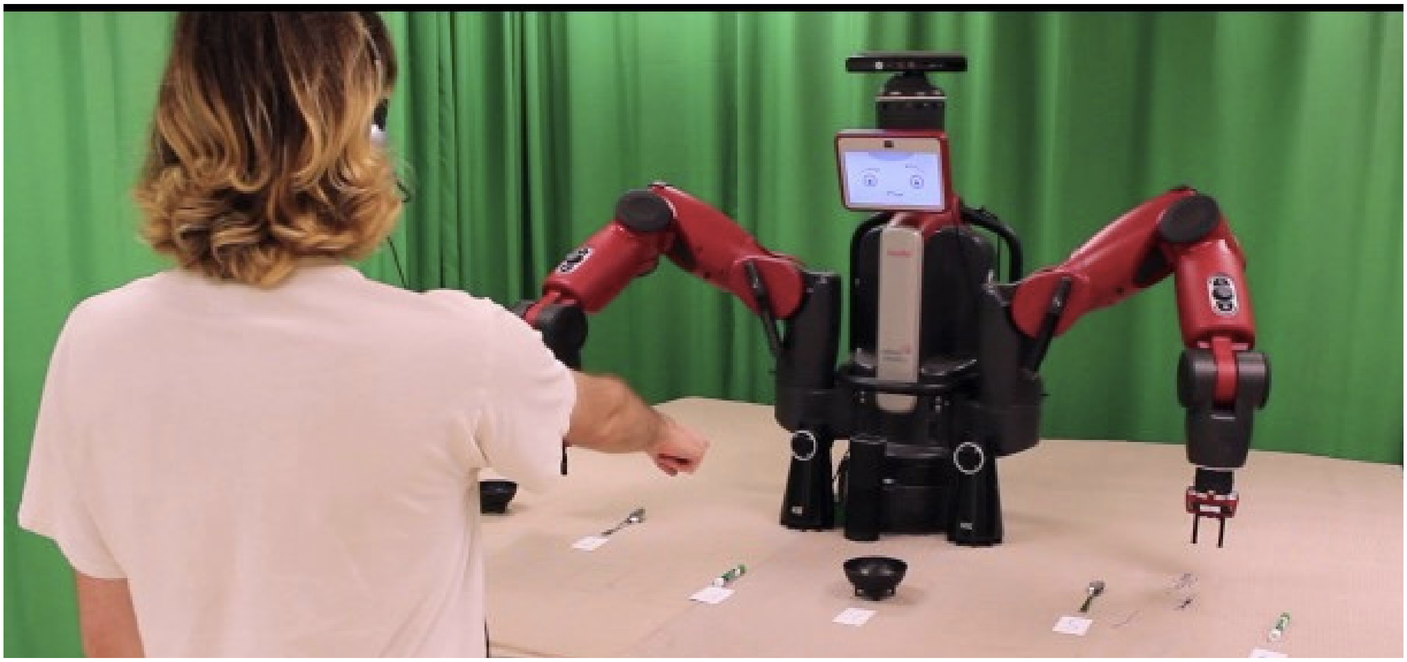


Figure 2: A robot takes gesture and voice input to modify it's Partially Observable Markov Decision Process "belief state" about objects with a human teacher. We hope to integrate this work into training robots to understand Moral Norms.

evaluates hypothesized reward functions through a form of counterfactual reasoning. For example, a community norm for a certain bodily distance can be detected by observing that this distance is maintained in spite of opportunities to select other distances. We have carried out preliminary experiments that show that learners can assign useful reward values to such norm preferences from observations of behavior. These experiments were carried out in toy simulated environments and we will assess in the proposed project whether and how these ideas scale to more realistic scenarios and eventually human-robot interactions.

In collaboration with Matthias Scheutz's human-robot interaction lab at Tufts University (https://hrilab.tufts.edu/), we have developed other formal approaches to norm learning (Sarathy, Scheutz, & Malle, 2017; Sarathy, Scheutz, Kenett, Allaham, Austerweil, & Malle, 2017). These approaches make use of the Dempster-Shafer (D-S) framework that models belief formation from uncertain and incomplete evidence. One advantage of the D-S framework is that it may be able to incorporate information gathered from community members' responses to norm violations. Across a number of evaluation studies we will compare the D-S and the IRL approach and develop potential hybrids that cover a broad range of learning situations.

Once an AIS has acquired norms that are reliably activated in their appropriate contexts, the AIS has taken the first step toward norm competence. And additional element is the ability to communicate its norm awareness and readiness to act in accordance with these norms. We believe that designing machines with such norm communication competence can make human-machine interaction vastly more beneficial. In a series of cognitive and interaction experiments (some involving augmented reality settings currently developed at HCRI), we will measure four metrics of impact of norm competence on human-machine interactions:

- humans will better understand norm-guided machine behavior, because it will be familiar, often well-adjusted to people's own behavior (this benefit connects directly to our research program on AIS explainability);
- norms offer mutual predictability through a limited repertoire of acceptable and likely actions;
- machines with norm competence will instill justifiable human trust, as people will feel safe and comfortable interacting with agents that act on shared norms and values; and
- as a result of this predictability, intelligibility, and trust, team actions will be better coordinated and more effective and group cohesion will increase.

Psychology deliverables:

1. A series of experiments that document the principles of norm learning and how they interact with different types of norms (prohibitions, prescriptions) and different levels of norm strength (high vs. low community demand).
2. A series of experiments examining whether AIS (especially robots) with norm competence perform better in select collaboration tasks than those without such competence. Experiments (online as well as in augmented reality) will measure objective behavior predictability and consistency; human understanding, acceptance, and trust; and objective task effectiveness.

RESEARCH DETAIL 2: MORAL NORMS

CS deliverables:

1. Algorithms for norm learning that reflect at least some of the principles of norm learning identified in the human experiments. Exploration of different formal approaches such as IRL and, in collaboration with Tufts computer science, D-S.
2. Creating a setup of augmented reality experiments that bridge robot control operation systems (e.g., ROS) with the augmented reality displays (e.g., Holens) to create realistic situations that can also be implemented in real robots.

References

Freeman, L. C. (1978). Centrality in social networks: Conceptual clarification. Social Networks, 1, 215–239.

Malle, B. F., Scheutz, M., & Austerweil, J. L. (2017). Networks of social and moral norms in human and robot agents. In M. I. Aldinhas Ferreira, J. Silva Sequeira, M. O. Tokhi, E. E. Kadar, & G. S. Virk (Eds.), A World with Robots: International Conference on Robot Ethics: ICRE 2015 (pp. 3–17). Cham, Switzerland: Springer International Publishing.

Newman, M. (2010). Networks: An introduction. New York: Oxford University Press.

Sarathy, V., Scheutz, M., Kenett, Y., Allaham, M. M., Austerweil, J. L., & Malle, B. F. (2017). Mental representations and computational modeling of context-specific human norm systems. In G. Gunzelmann, A., Howes, T. Tenbrink, & E. J. Davelaar (Eds.), Proceedings of the 39th Annual Conference of the Cognitive Science Society (pp. 1035-1040). Austin, TX: Cognitive Science Society.

Sarathy, V., Scheutz, M., & Malle, B. F. (2017). Learning behavioral norms in uncertain and changing contexts. Proceedings of the 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom). Debrecen, Hungary.