Enhancing Agent Safety through Autonomous Environment Adaptation

Benjamin Rosman
Mobile Intelligent Autonomous Systems
CSIR
South Africa
brosman@csir.co.za

Bradley Hayes
Yale University
51 Prospect St #504
New Haven, CT, USA
bradley.h.hayes@yale.edu

Brian Scassellati
Yale University
51 Prospect St #501
New Haven, CT, USA
scaz@cs.yale.edu

***Abstract*—Exploration and self-directed learning are valuable components of early childhood development. This often comes at an unacceptable safety trade-off, as infants and toddlers are especially at risk from environmental hazards that may fundamentally limit their ability to interact with and explore their environments. In this work we address this risk through the incorporation of a caregiver robot, and present a model allowing it to autonomously adapt its environment to minimize danger for other (novice) agents in its vicinity. Through an approach focusing on action prediction strategies for agents with unknown goals, we create a model capable of using expert demonstrations to learn typical behaviors for a multitude of tasks. We then apply this model to predict likely agent behaviors and identify regions of risk within this action space. Our contribution uses this information to prioritize and execute risk mitigating behaviors, manipulating and adapting the environment to minimize the potential harm the novice is likely to encounter. We conclude with an evaluation using multiple agents of varying goal-directedness, comparing agents' self-interested performance in scenarios with and without the assistance of a caregiver incorporating our model. Our experiments yield promising results, with assisted agents incurring less damage, interacting longer, and exploring their environments more completely than unassisted agents.**

I. Introduction

Caregiver-guided and free exploration activities are known to be an important part of development. It has been shown that the interpretation of goal-directed spatial behavior as intentional action can begin as early as 12 months [1], suggesting that caregiver behaviors may impart high-level information to infants. The same study shows 12-month-olds to be even capable of evaluating an action's rationality in limited circumstances, indicating a preliminary understanding of goal-directed behaviors. Particularly for children between one and two years of age, exploratory activity is critical to development [2].

Earlier work in examining exploratory behaviors of infants and toddlers has shown there to be direct correlations between these behaviors and problem-solving ability [3]. At 12 months of age, studies indicate that a greater breadth of exploratory behavior, when faced with novel objects (toys), was linked to both an increased quantity of behavior as well as more successful and sophisticated problem-solving ability. These connections continue into toddler years, as others [2] have shown that exploratory style at 19 months of age is predictive of typical vs. delayed developmental level (measured by pretend play level and performance of meaningful actions) at 30 months, painting a more complete picture of developing competence (effective adaptation).

Two-year-olds inherently include social interactions in their exploratory behaviors, learning about the world around them. For example, the presence of a toddler's mother during a novel exploration task is known to elicit social behaviors, creating active and passive bids for involvement and support of action during exploration. Children that make more bids to their mothers during exploration do so more actively than those that do not [4], [5]. This suggests that a trusted caregiver may be able to encourage deeper and more plentiful interactions.

Toddlers commonly encounter issues where goal-directed exploration and learning is limited by the presence of environmental hazards. Situations can occur in which a caregiver might wish a toddler to learn to avoid a particular action (e.g., touching a hot stove-top) without the toddler experiencing the environmental punishment for doing so (e.g., receiving a painful burn). Another common scenario may involve discouraging a child from exploring or playing near stairs or other falling hazards, without him directly experiencing the potential negative consequences of interacting in such an area. The trauma from experiencing such hazardous events can have lasting developmental effects, persisting even through adulthood [6].

In this work, we propose a novel shaping mechanism by which a robot can reduce a novel agent's inherent risk during exploration within an environment. Our proposed method accomplishes this without sacrificing the quality of the learning experience, producing behavior for an assistive robot that can autonomously adapt a novice agent's environment to reduce the risk of damage to themselves or to objects nearby. To do so, this robot must monitor the behavior of the person, predict intended goals, and take non-invasive, preventative measures to ensure the exploring agent's safety.

We accomplish this without any direct communication, as all information is implicitly conveyed through manipulations to, and exploration within, the environment. This allows for

the application of our contribution to agents of unknown capabilities and with agents incapable of direct communication. As an example, if the person is at risk of collision with any objects, those objects should be protected accordingly (perhaps by placing a barrier around them or temporarily relocating them somewhere safer).

To this end, we extend previous work on how to autonomously give advice to an agent that may be lost, may be exploring its surroundings, or may have a suboptimal policy [7]. Drawing on these ideas, we present a model allowing an assistant or caregiver observing the environment to learn a model of normal behaviors from successful trajectories through the space. Deciding if a new agent moving through the environment needs assistance is then a form of anomaly detection, by determining how well that new agent fits the behavior predicted by the model.

Minimally invasive assistance to a novice agent may take the form of adaptive changes to the environment, such as moving a fragile or hazardous object to a less risk-prone area or adding a temporary barrier to that area. This indirect assistance is proposed in opposition to more traditional modes of direct assistance that use physical manipulation, where the assistant's action policy is imposed [8] (for instance, by being led by the hand through a task). More similar to our work are methods of advice giving with restricted frequency (such as [9], [10]), where a teacher can influence a learner's policy at some non-trivial action cost. Our approach characterizes the generation and evaluation of indirect assistance as a means of maximizing the safety of fellow agents in an environment while affording them similar levels of autonomy to unassisted scenarios.

Such indirect interventions are not without precedent in the robotics or machine learning literature. The benefit of interactive shaping, the practice of providing targeted feedback to manipulate portions of an agent's policy to facilitate goal achievement, is well established [11], [12], [13]. While the ability of an agent to interpret and respond to live feedback or to freely explore a task space can greatly improve the convergence rate of its action policy, there exist many cases where catastrophic failures may occur. In our chosen domain, these failures are unacceptable to use as learning opportunities as they may result in disasterous injury.

Accordingly, we model a novice agent's risk with the expectation of perturbation from typically observed behaviors derived from expert demonstrations. Our approach accomplishes this by probabilistically merging collections of task-based risk models, and choosing environmental manipulations in order to minimize the expected risk given the potential execution policies of the agent.

The lack of direct communication between novice and caregiver introduces the requirement of estimating the probability that an agent is behaving safely based on its trajectory. The assistant may then adapt the environment to mitigate the risks inherent to the nearest danger zones. We demonstrate the utility of our approach by evaluating its effectiveness at providing a safe interaction environment for agents with uncertain or suboptimal navigation policies and behaviors.

In the following section we introduce term definitions and necessary background information. We then describe our process of building models to assess safety and risk given collections of expert demonstrations for a variety of tasks. Finally, we conclude with an evaluation within a simulated online risk mitigation domain, validating our approach through various metrics related to safe exploration and goal achievement for agents of varying levels of goal-directedness.

II. Preliminaries

Throughout the remainder of this paper, we employ the following terminology. The *novice* is moving around the environment, the *caregiver* watches the behavior of the novice and adapts the environment, and *objects* are regions with which the novice can collide. Such a collision causes *damage*, which can be treated as a negative reward in the context of reinforcement learning. To prevent damage, the caregiver manipulates the environment such that the novice still negatively reinforces undesirable behaviors but without incurring the full damage that such a behavior may have otherwise inflicted.

Let an environment be specified by a set of states S, and the actions available to an agent are drawn from a set A. An agent in a state $s \in S$ can select an action $a \in A$, after which the agent transitions to another state $s' \in S$ according to a probability distribution given by the transition function $T(s, a, s')$. This transition results in a reward given by the reward function $R(s, a)$. The motion of an agent through the environment is thus given by the Markov decision process (MDP) (S, A, T, R) [14].

A decision rule for probabilistically (or deterministically) selecting actions as a function of state is a policy $\pi : S \times A \longrightarrow [0, 1]$. An optimal policy π^* is the policy which maximizes total reward from every state of the MDP.

Action priors [15] are a recently proposed mechanism for learning models of general behavior in a common domain across multiple tasks. These involve maintaining distributions over the action set for each state, corresponding to the number of known optimal policies that select each action at that state. Particularly for domains in which goals cannot be or are not fully specified, we include in our set of optimal policies the set of 'best known' policies for a given task. This provides a basis for learning goals from inverse reinforcement learning [16] approaches, where expert demonstrations can train a reward function in lieu of a specified goal state descriptor.

For each state s in the environment, the action prior $\theta_s(a)$ is computed from $\alpha_s(a)$, which is the number of optimal policies (or trajectories) π in the set of all such policies Π in which a is a reasonable action choice in s (i.e. a is taken in s with probability greater than some threshold δ). Thus,

$$\alpha_s(a) \quad = \quad \|\{\pi \in \Pi | \pi(s, a) > \delta\}\| + \alpha_s^0(a),$$

where $\|\{\cdot\}\|$ represents the size of a set, and $\alpha_s^0(a)$ is a hyperprior used to prevent overfitting [15], by allowing for that fact that the training policy set Π may not be fully representative of the set of all behaviors in the environment.

The action priors are then typically computed as a draw from a Dirichlet distribution: $\theta_s(A) \sim \mathrm{Dir}(\alpha_s(A))$. This state-based distribution over the action set assigns probability to each action based on the number of tasks for which that state-action combination was optimal. This therefore provides a model of optimal behavior in the environment, aggregated across all known tasks.

III. A Model of Safety

We approach the problem of facilitating safe learning and exploration initially as that of anomaly detection. Given a model of safe interaction, described by the set of MDP reward functions within known tasks and encapsulated as action priors, we compare observed agent behaviors to these known policies. This model construction is particularly robust as it makes determinations independent of an agent's goals, perceptual capabilities, or assumed knowledge of the world. By using such an approach, we also avoid the pitfall of relying on constructing models of unsafe interaction, which may not be feasible or reasonable to construct.

We would like to compute the probability that the observed trajectory was drawn from a model of abnormal or risky behavior, but to learn such a model one would need examples of this (which is undesirable from the point of view of both the agent and the environment). Instead we compute the probability of the trajectory having been drawn from a model of normal behavior, $P(trajectory|model)$, which is something that can be learned from experience.

Our model requires computing the probability that a novice agent is behaving safely. Let the trajectory taken by the novice thus far be represented by a time-indexed alternating state and action sequence, as

$$\tau \quad = \quad s^{t+1}, a^t, s^t, a^{t-1}, s^{t-1}, \ldots.$$

We compute the probability that a novice is behaving safely by the probability that the trajectory followed by the novice was drawn from a model of normal motion in the environment, as modelled by the action priors. By Bayes' rule, this is given by

$$P(safe|\tau) \quad = \quad \frac{P(\tau|safe)P(safe)}{P(\tau)} \tag{1}$$

$$= \frac{\prod_{k=1}^{t} \theta_{s^k}(a^k) P(safe)}{\prod_{k=1}^{t} \theta_{s^k}(a^k) P(safe) + \prod_{k=1}^{t} \rho_{s^k}(a^k)(1 - P(safe))}$$

where $P(safe)$ is the prior belief that the agent is choosing a safe action at each timestep, and $\rho_{s^k}(a^k)$ is an action distribution for an unsafe agent.

As we do not know *a priori* the model of unsafe behavior and are unable to collect this data, we choose to represent this model as one where every action is assumed to occur with equal probability. A richer model of goal likelihoods could improve upon this, however doing so imposes requirements of deep domain knowledge and rich observation on the caregiver to collect or use such data. As such, in our experiments we let $\rho_s(a) = 1/|A|$, $\forall s \in S, a \in A$. Furthermore, without *a priori* knowledge, we assume a uniform prior over the probability of the agent behaving safely, and let $P(safe) = 0.5$.

IV. A Model of Danger

Having computed a probabilistic estimate of how unsafe the behavior of a novice agent is given a model of expert motion, the caregiver is required to modify the environment to enhance the safety of the agent.

Damage is caused by a collision between an agent and any of a number of objects in the environment. The caregiver must therefore estimate which object poses the greatest risk to the novice. To do so, it computes the expected damage caused by a collision with each object o. This is determined as the damage that would be caused by a collision between the novice and that object, weighted by the probability of that collision, as

$$\begin{aligned} \mathbb{E}(d_o|\tau) &= P(collision|\tau) \times d_o \\ &= (1 - P(safe|\tau)) \times P(reach_o|\tau) \times d_o \end{aligned} \tag{2}$$

where $\mathbb{E}(d_o|\tau)$ is the expected damage that could be caused by a collision with object o given the current trajectory τ of the novice, $P(safe|\tau)$ is the probability of the novice behaving safely as given by Equation (1), $P(reach_o|\tau)$ is the probability the novice reaches and collides with o, and d_o is the extrinsic cost of the damage caused by such a collision. As a proxy for $P(reach_o|\tau)$, we represent this quantity by the normalized distance of the agent from o. This provides an estimate of the number of timesteps that would be required for the novice agent to reach and collide with the object from its current position.

The damage cost d_o is extrinsically defined, and associated with different items in the environment based on the danger that a collision with an agent may pose to either the agent or the object. We compute an expected intrinsic cost to the agent, based on the distance of the agent from that object, as well as the estimated probability of a collision.

The caregiver utilizes the following action policy: at each timestep, compute the expected harm posed by each object using Equation (2), and select the object o maximizing potential harm. The risk mitigation strategy then depends on the class of o. In our experiments, for example, if o is stairs then the robot moves to block them, whereas if o is a candle then the robot will pick this up and move it to a safe table.

V. Experiments

We validate our approach in an experiment simulating a toddler (novice agent) exploring a household domain. In each goal-directed episode the novice agent is attempting to navigate to a sequence of randomly selected toy bins within the environment. In addition to walls and toy bins, this environment contains several hazards, some mobile and others immobile. Hazards which may not be moved include stairwells and tables, while those that can be relocated are candles that sit upon tables. We explore the behavior of our caregiver robot by examining its interactions with different novice agent types, varying the novice's goal-directedness via exploration likelihood, as they interact with the environment

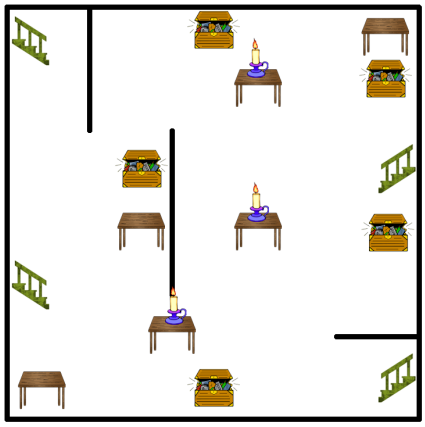

Fig. 1. The map used in the experiments. The environment consists of open spaces, walls, staircases, tables, candles, and toy bins.

and perform various navigational tasks (in the form of toy retrievals).

A. Environment

Our experiment utilizes an environment characterized by a house's floor plan, discretized into a 2D grid world. Apart from walls and free space, the environment has five types of objects: candles on tables which may start fires if collided with (major damage), walls, tables which may be bumped into (minor damage), stairs that the agent may fall down (major damage), toy bins that act as goal locations for the novice, and the mobile caregiver robot. The domestic environment used in the experiments is shown in Figure 1.

B. Novice Agent Behavior

Novice agents are initialized at a random start location with knowledge of all toy bin (goal) locations, though the rest of the environment is unknown. Without a goal, a novice will randomly choose a goal destination from the list of known toy bins. When a goal has been selected, the novice agent will follow an $\epsilon-$greedy policy to achieve it before selecting a new goal, meaning it will take an optimal action $\epsilon\%$ of the time and a random action $(100-\epsilon)\%$ of the time. In our simulation, we allow novices to 'play' for a maximum of 200 timesteps, with each timestep corresponding to the time required to move one grid square in the environment. The action set for these agents is limited to 4-connected movement within the environment. Collisions with objects or walls result in the agent staying in place and incurring the appropriate penalty. The amount of damage is extrinsically defined as being 5 for a collision with a stairwell, 4 for a collision with a candle, and 1 for a collision with an empty table. Collisions with walls or toy bins do not incur any damage.

C. Caregiver Training and Behavior

The model of normalcy we utilize is developed from expert trajectories of an optimal agent moving through the room. The expert trajectories take the form of collision-free, shortest-path routes from random start locations to each of the specified goal locations (toy bins), for a total of $1,000$ timesteps. These optimal policies provide the action priors required to inform the risk mitigation strategy of the caregiver. Our evaluation simulates a caregiver robot that executes up to 3 actions for each 1 of the agent, assuming a movement speed ratio similar to that of healthy adult humans (5.0 km/h) [17] to healthy 1-2 year olds (1.6km/h) [18]. The action set for the caregiver consists of 7 actions: 4-connected grid movement operations, a wait action, a get-object action, and a place-object action.

There are two intervention strategies which can be employed by the caregiver. Firstly, it is able to pick a candle up off a table, and move it to another. Secondly, the caregiver may wait in a stairwell, which blocks the toddler from accessing it and subsequently incurring a large amount of damage.

D. Evaluation Criteria

The performance of our algorithm is characterized by the average damage incurred by a novice agent over the course of its interactions (Figure 2 and Table I), the number of timesteps elapsed before the novice reached a critical damage threshold that terminates the simulation episode (Figure 3 and Table II), and the amount of environmental coverage the agent was able to achieve during exploration (Figure 4). As the goal of our algorithm is to actively mitigate and minimize the risks inherent to interaction and exploration in a novel environment, the average incurred harm is an essential metric to track. We are also interested in examining how much time the novice is able to use for exploration/task execution in the environment before it is forced to stop as a result of accumulating too much damage. This is related to our final metric, the amount of environment coverage that the novice achieved (equivalent to environmental knowledge gained) that can be used to better inform its navigation and other relevant interactions in that space.

The results reported in this section are all averaged over 20 runs (episodes) each of the novice agent interacting in the environment both with and without the caregiver robot present. As the values we present will change with different environments, objects, and damage definitions, we identify the trends of the results as being more important than the particular values themselves.

E. Results

Our results show definitive and tangible benefits achieved by the risk mitigation algorithms proposed within our work. Novice agents with caregivers present received less damage from the environment (Figure 2), spent more time interacting (Figure 3), and explored more of the environment before receiving a critical amount of damage (Figure 4).

In Figure 2, we evaluate the average damage incurred by the novice over each of its interaction episodes, lasting 200 timesteps. It is expected that an agent with a higher exploration rate will encounter more harmful areas of the environment, and as such experience more damage. The inclusion of our caregiver agent generally reduces this incurred damage by over 66%, performing mitigation strategies informed by known optimal behaviors and predicted deviations.

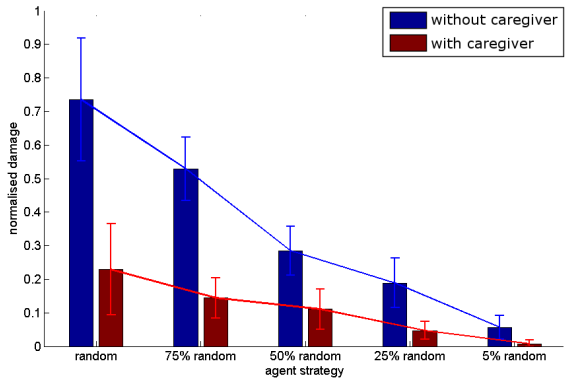

Fig. 2. The average normalized damage (negative reward) accumulated by novice agents of various levels of goal-directedness, both with and without the assistance of a caregiver robot. Error bars represent one standard deviation from the mean. A value of 1.0 on this graph corresponds to the maximum damage received in a single episode across all agent types and conditions. Novices followed an ϵ-greedy policy to their goal, determining the frequency of choosing optimal or random exploratory actions.

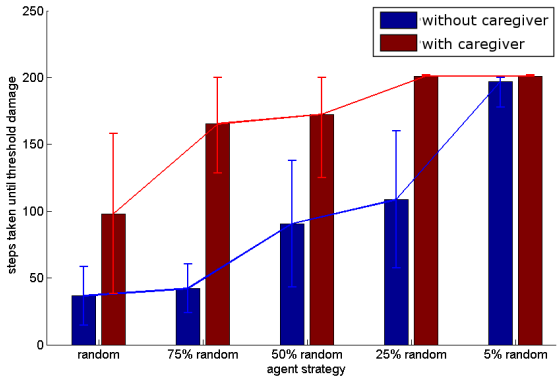

Fig. 3. The average number of timesteps (maximum 200) before the novice agent incurred damage above the stoppage threshold. Error bars represent one standard deviation from the mean. Values of 200 indicate that the agent completed the entire interaction without incurring damage above the termination threshold.

Agent	No caregiver	With caregiver
random motion	0.7346	0.2291
75% random	0.5295	0.1438
50% random	0.2846	0.1106
25% random	0.1887	0.0466
5% random	0.0565	0.0072

TABLE I
MEAN NORMALIZED DAMAGE (NEGATIVE REWARD) ACCRUED BY VARIOUS NOVICE AGENTS WITH AND WITHOUT CAREGIVER INTERVENTION

Agent	No caregiver	With caregiver
random motion	35.50	97.05
75% random	41.05	164.20
50% random	89.50	171.50
25% random	107.70	200.00
5% random	195.75	200.00

TABLE II
AVERAGE NUMBER OF TIMESTEPS UNTIL THRESHOLD DAMAGE WITH AND WITHOUT CAREGIVER INTERVENTION

While each episode lasted 200 timesteps, it was also recorded when an agent sustained above a critical threshold of damage (corresponding to 15, or 3 times the damage of a staircase collision). We examine the number of timesteps the agent is able to complete, before either sustaining too much damage to continue or reaching the end of the episode (fixed at 200 steps) in Figure 3. Understandably, less goal-directed novice agents are more prone to encountering harmful objects in the environment. On average, fully random and 75%-random agents without caregiver assistance do not even complete a quarter of the episode before sustaining critical damage. Unassisted agents with ϵ-greedy strategies of 50% and 25% often only complete half of the episode. Even with a 5% exploration rate, the agent in the non-caregiver condition does not always complete the episode before sustaining critical damage, illustrating the danger inherent within the test scenario environment.

When we introduce and train our proposed caregiver agent, the number of timesteps before sustaining critical damage improves substantially. The caregiver mitigates risk even in the fully random novice, an agent entirely without goal directed motion, by doubling the number of actions taken prior to episode termination. Goal directed novices experience even more dramatic improvements, with 75%-random and 50%-random agents completing over 150 of the 200 possible timesteps on average, with many of them completing the entire episode without receiving critical damage. For 25%-random and 5%-random agents the caregiver robot was able to ensure *every episode* ran to completion, with no agents having their interaction terminated early. These results are indicative of the clear benefits afforded by the caregiver's control policy, achieved through accurate safety prioritization and action prediction.

Finally, we examined the overall environment coverage achieved by the various novices prior to reaching the damage threshold (Figure 4). It is fairly straightforward to expect that a weakly goal-directed agent (one with a high exploration rate) that is able to spend more time interacting in the environment will likely cover more ground than a strongly goal directed or temporally limited agent. Accordingly, we observe coverage that varies in proportion to both the exploration rate and episode duration.

In the non-caregiver condition, the entirely random novice agent covers fewer than 15 cells on average, with the 75%-random agent not faring much better. The 50%-random agents, using nearly double the actions on average to explore before sustaining critical damage, achieve approximately double the environment coverage as their less directed counterparts. Finally, the 25%-random and 5%-random agents perform best, reaching nearly 35 states on average.

When interacting in an environment with the caregiver robot, the novices generally outperform their solo counterparts with only the 5%-random agent having similar results (largely due to its limited deviation from an optimal trajectory). This coverage increase is directly attributable to the increased duration of exploration available to the agent. We include these

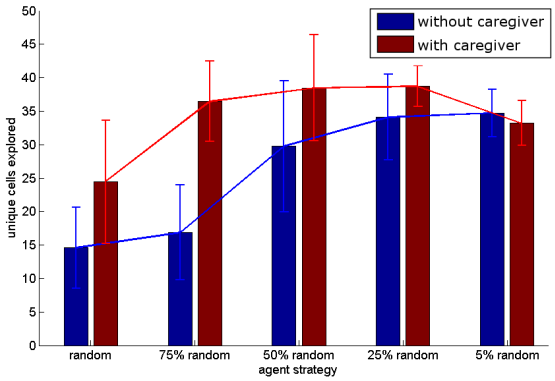

Fig. 4. The average number of unique grid cells the novice visited per episode. Error bars represent one standard deviation from the mean. The number of cells visited is used as a metric to evaluate the amount of environmental knowledge gained by the novice agent over the course of an episode (up to 200 timesteps).

results to provide a concrete measure of the realized benefit from having the additional time for environmental exploration and interaction. We show that not only does the caregiver robot provide a safer environment, but this risk mitigation results in actual increases in environmental exploration, and subsequently potential increases in the agent's experience diversity and its awareness of its surroundings. Accordingly, our approach showcases the benefits of caregiver scaffolding relating to improving an agent's self-sufficiency and task proficiency.

VI. Discussion and Conclusion

We present a model usable by a caregiver agent to assist novice agents through the reduction of danger inherent in an environment. We accomplish this through the utilization of expert demonstrations to learn regular behaviors for multiple tasks within an environment. From this data we create a model of normal behaviors, using it to enable an autonomous agent to mitigate the highest-risk scenarios the agent it's assisting is predicted to encounter. In doing so we show how an environment can be adapted online to respond to the assisted agent's behaviors and goals in the absence of directly communicated information specific to the agent's intent.

Our results show that our approach is effective in reducing harm experienced by a novice agent interacting in the same environment as the caregiver. This manifested other important benefits, including increased interaction and exploration time (prior to receiving a critical amount of damage) as well as increased coverage of the environment. The trends within our data demonstrate the value of such a caregiver, particularly due to its lack of reliance on direct communication or knowledge about a specific agent. Our approach can be extended to partial or mixed observability domains where the current state of the environment is not reliably, globally known to the caregiver by modeling potential harm over policies extending from a particular belief state of the world. Developing robust exploration methods for this domain, characterizing the tradeoffs between taking assistive actions and performing environmental observation updates remains as future work.

Providing a means of autonomously adapting an environment for a novice agent to make exploration less risk-prone is valuable within many contexts. In addition to mitigating risk for developing humans, the same approach can be used to assist robots learning tasks within an environment. The same benefits persist, as a robot that is able to perform more iterations of an action or explore the action space more completely without damaging itself would be expected to achieve better task proficiency than one without these advantages.

In future work we wish to explore different ways in which the assistant may help an agent perform a task on a more general class of problems, generalizing known expert policies to novel tasks, environments, and agent configurations.

References

[1] G. Gergely, Z. Nádasdy, G. Csibra, and S. Biro, "Taking the intentional stance at 12 months of age," *Cognition*, vol. 56, no. 2, pp. 165–193, 1995.

[2] A. S. Rusher, D. R. Cross, and A. M. Ware, "Infant and toddler play: Assessment of exploratory style and development level," *Early Childhood Research Quarterly*, vol. 10, no. 3, pp. 297–315, 1995.

[3] D. A. Caruso, "Dimensions of quality in infants' exploratory behavior: Relationships to problem-solving ability," *Infant Behavior and Development*, vol. 16, no. 4, pp. 441–454, 1993.

[4] L. C. Mayes, A. S. Carter, and D. Stubbe, "Individual differences in exploratory behavior in the second year of life," *Infant Behavior and Development*, vol. 16, no. 3, pp. 269–284, 1993.

[5] K. Pridham, P. Becker, and R. Brown, "Effects of infant and caregiving conditions on an infants focused exploration of toys," *Journal of Advanced Nursing*, vol. 31, no. 6, pp. 1439–1448, 2000.

[6] B. D. Perry, "Neurobiological sequelae of childhood trauma: Ptsd in children." pp. 253–276, 1994.

[7] B. S. Rosman and S. Ramamoorthy, "Giving advice to agents with hidden goals," in *IEEE International Conference on Robotics and Automation*. IEEE, 2014, pp. 1959–1964.

[8] B. Akgun, M. Cakmak, K. Jiang, and A. L. Thomaz, "Keyframe-based learning from demonstration," *International Journal of Social Robotics*, vol. 4, no. 4, pp. 343–355, 2012.

[9] L. Torrey and M. Taylor, "Teaching on a budget: Agents advising agents in reinforcement learning," in *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2013, pp. 1053–1060.

[10] B. Hayes and B. Scassellati, "Discovering task constraints through observation and active learning," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014, pp. 4442–4449.

[11] T. Erez and W. D. Smart, "What does Shaping Mean for Computational Reinforcement Learning?" *International Conference on Development and Learning*, pp. 215–219, 2008.

[12] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum Learning," *International Conference on Machine Learning*, 2009.

[13] W. B. Knox and P. Stone, "Interactively Shaping Agents via Human Reinforcement: The TAMER Framework," *International Conference on Knowledge Capture*, September 2009.

[14] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. The MIT Press, 1998.

[15] B. S. Rosman and S. Ramamoorthy, "What good are actions? Accelerating learning using learned action priors," *International Conference on Development and Learning and Epigenetic Robotics*, November 2012.

[16] A. Y. Ng, S. J. Russell *et al.*, "Algorithms for inverse reinforcement learning." in *Icml*, 2000, pp. 663–670.

[17] R. C. Browning, E. A. Baker, J. A. Herron, and R. Kram, "Effects of obesity and sex on the energetic cost and preferred speed of walking," *Journal of Applied Physiology*, vol. 100, no. 2, pp. 390–398, 2006.

[18] G. Cavagna, P. Franzetti, and T. Fuchimoto, "The mechanics of walking in children." *The Journal of Physiology*, vol. 343, no. 1, pp. 323–339, 1983.