Safety in AI-HRI: Challenges Complementing User Experience Quality

Richard G. Freedman and **Shlomo Zilberstein**
College of Information and Computer Sciences
University of Massachusetts Amherst
{freedman, shlomo}@cs.umass.edu

Abstract

Contemporary research in human-robot interaction (HRI) predominantly focuses on the user's experience while controlling a robot. However, with the increased deployment of artificial intelligence (AI) techniques, robots are quickly becoming more autonomous in both academic and industrial experimental settings. In addition to improving the user's interactive experience with AI-operated robots through personalization, dialogue, emotions, and dynamic behavior, there is also a growing need to consider the safety of the interaction. AI may not account for the user's less likely responses, making it possible for an unaware user to be injured by the robot if they have a collision. Issues of trust and acceptance may also come into play if users cannot always understand the robot's thought process, creating a potential for emotional harm. We identify challenges that will need to be addressed in safe AI-HRI and provide an overview of approaches to consider for them, many stemming from the contemporary research.

1 Introduction

The presence of robots in the real-world is slowly becoming a reality, and an upcoming step in this transition is the direct interaction between humans and robots. Furthermore, many present-day robots contain artificial intelligence (AI) to provide more dynamic and realistic interactive experiences with the users. In such cases, many robots that have previously been allowed to directly interact with humans are quite fragile like children's toys and used for mostly non-physical interactions such as communications (Kory and Breazeal 2014), presentations (Knight and Simmons 2013; Hoffman and Weinberg 2010), and playing games (Hirose, Hirokawa, and Suzuki 2014). This design for a robot is not practical for more physical interactions, which are needed in many domains such as industrial factories (Levine and Williams 2014; Lasota, Rossano, and Shah 2014), furniture assembly/moving (Mörtl et al. 2012), and search and rescue (Nourbakhsh et al. 2005) — these robots have often been constrained to work in separate areas from humans to avoid physical safety concerns (Enright and Wurman 2011). Some cases of interaction even involve issues for mental safety due to potential emotional bonds/relationships (Scheutz and Arnold 2016; Darling, Nandy, and Breazeal 2015) and uncensored content (Price 2016).

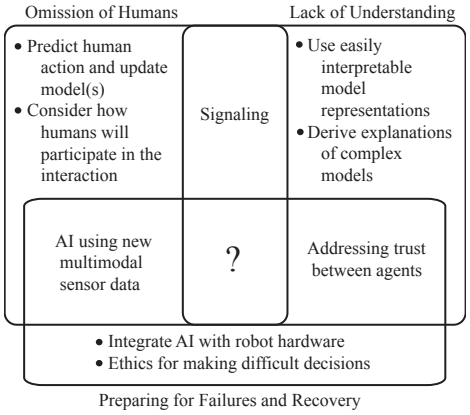

Figure 1: Proposed approaches to categories of safety in AI-HRI. We are unaware of any that handle all categories (marked by ?).

Developing safe technologies for AI and human-robot interaction (HRI) has been discussed in the past for specific algorithms and applications, but there are generalizations that can be proposed and addressed. The recent SafePlan (Shafti et al. 2016) and SāfArtInt (OSTP and CMU 2016) Workshops are introducing this area, and our goal is to extend its reach to the AI-HRI community. Many existing works do not mention their potential applicability towards safety in AI-HRI, but *techniques for improving the user experience with robots has a lot in common with ensuring their safety*. In fact, the relationship is bidirectional as *ensuring safety also improves the user experience*. In the following sections, we present general categories of challenges for safety in AI-HRI and propose approaches, summarized in Figure 1. Based on current literature in AI-HRI, we also explain how these approaches can start to be addressed. We then conclude with an overall summary and consider future directions.

2 Omission of Humans

In AI, many domain instances of automated problem solving and machine learning (ML) represent the environment and autonomous agent(s) as the only components of the problem. However, humans are not often encoded in these environments due to their unpredictability, even with more accurate mental models. A robot planning to move between rooms or perform a tactile task has enough difficulties considering the geometrical space around itself and its own configura-

tion space; patternless moving obstacles that may or may not interfere are not regarded as a priority for many laboratory settings. However, if we consider the robots in the real world, then we begin to observe that *humans are walking around everywhere in these environments and have many interactions with the autonomous machine*, both positive and negative (Yang et al. 2015; Reben and Paradiso 2011).

Ignoring the human obstacles during planning and training a policy/function can be unsafe with heavier and stronger robots that may injure people by simply running over their foot, charging into them (McFarland 2016), or accidentally hitting them with a swinging appendage. Simply detecting an obstacle before progressing with each action, though effective, is not efficient. Actions have potentially long-term consequences such as entering a long hallway from one end while humans progress from the other end. The robot now becomes as much an obstacle as the humans, and one of them must backtrack to allow the other(s) to exit. Clearly this is unavoidable if such a hallway is the only path that connects two areas of a building, and the case of humans using this hallway to escape from an emergency situation (fire, chemical spill, etc.) is hopefully not a common scenario in daily applications. However, the application of safety is not an excuse to develop systems that act less optimally in the average case. Unhelkar et al. (2015) recently introduced a method for predicting a user's trajectory to adapt the path planning space, altering the paths the robot will consider taking in the near future to avoid unnecessary confrontation. When confrontations cannot be avoided in tight spaces, Mainprice, Rafi, and Berenson (2015) introduce the use of predicting occupancy regions for subtasks that the human partner will approach, and then the robot can deal with subtasks in other regions that will not interfere with the human.

By dividing up the tasks and avoiding confrontation in the works above, the robot is able to reduce the likelihood of directly working with the human so that it may omit the human afterwards during its task execution. Some tasks must involve the use of humans to assist the robot when it is incapable of performing a specific action for its task, such as repositioning misplaced/fallen objects (Knepper et al. 2015) or pushing an elevator button due to the lack of limbs (Rosenthal and Veloso 2012), or when there are shared resources that require cooperation between the humans and the robots to each use them. In these situations, signaling is important to inform the human of the robot's own intent — communication can improve safety by enabling the human, a far more robust and dynamic planner than current state-of-the-art AI, to accommodate rather than get in the way. The human's greater performance at such tasks has inspired the use of imitation learning to train probabilistic motion primitives for interaction (Maeda et al. 2014) so that the robot is able to react to human movements in ways that mimic the observed reacting human. A slightly less optimal motion-planning trajectory can also display a robot's intent for which resource it wants from a collection by emphasizing a position disambiguating between them (Dragan and Srinivasa 2014), and a robot can perform supportive actions that provide a human partner clues indicating suggestions for their shared task (Hayes and Scassellati 2015).

In addition to the mentioned works, research in semi-autonomous systems (Zilberstein 2015) investigates tasks where the AI agent must develop solutions that require some level of human involvement in order to guarantee completion. Conversely, Levine and Williams (2014) and Freedman and Fukunaga (2015) propose the integration of plan recognition and planning so that agents can identify a human's task and act with respect to this task. The primary difference between these approaches is that the semiautonomous robot is the one assigned the task (where humans assist when needed) while the integrated recognition and planning system assumes that the human is the one assigned the task (where the robot assists as much as it can).

3 Lack of Understanding

While Section 2 represents the ignorance of people as a robot not understanding humans, the challenge applies equally in the other direction. Not only does a robot have difficulty accounting for humans in its problem solving algorithms, but *humans often have difficulty interpreting robot behaviors*. Many ML methods learn functions that are not easily interpretable to humans such as the currently popular deep learning neural networks (Nguyen, Yosinski, and Clune 2015), and many of these functions and learned plans/policies for solving real-world problems are far too large and complex for a human to read. The humans' inability to understand the robots with which they are interacting is an unsafe practice because, just as a robot without consideration of others can harm a human by moving into her, a misunderstanding human can incorrectly assume she will not have contact with the robot. Additionally, a robot can learn actions for certain states that are harmful without an easy way to monitor it, such as making a mess in order to later clean it up for additional reward (Amodei et al. 2016).

This is likely one reason that many members of the AI-HRI community use hierarchical task networks (Erol, Hendler, and Nau 1994). Their hierarchical nature breaks down more complex tasks into simpler ones, and this top-down explanation is intuitive for humans to interpret. Although other representations are typically less intuitive, research has been done to develop descriptions or summaries that may facilitate a human's understanding. For complex policies with many states, compact contingency plan representations can sacrifice some optimality to summarize the main state-action pairs (Horstmann and Zilberstein 2003). Learned clusters can also be explained using prototype examples (Kim, Rudin, and Shah 2014) or features describing the 'average objects' (Freedman and Zilberstein 2016).

Although being on the same page as the robot can improve the likelihood of safety in interactions, there is still a safety concern if the human misinterprets the robot's status. For example, knowing that the robot will execute action a in state s does not help the user prepare unless she knows the robot is in s. This returns to the use of signals to provide internal information rather than just intent. Baraka, Paiva, and Veloso (2015) attached lights of varying colors to a Cobot in order to convey internal states to nearby people who may try to interact with it. There are also direct communication methods including announcements of takeover

in semiautonomous driving scenarios (Miller et al. 2015) and queries to confirm an understanding of the human's intentions (Mirsky and Gal 2016) or teachings (Cakmak and Thomaz 2012). Public signals not only improve communication, but also improve emotional safety by putting the humans at ease of better understanding what the robot is up to. However, private communication between multiple robots may be acceptable without negatively affecting the humans' trust (Williams et al. 2014). One future direction to consider for this safety challenge is signal protocols — Wizard-of-Oz experiments have revealed that the same signal can be interpreted differently per person (Sirkin et al. 2015) and that people may assume different signals to trigger a particular response (Pourmehr, Thomas, and Vaughan 2016).

4 Preparing for Failures and Recovery

Even when there is mutual understanding between agents and they are interacting based on a concensus, *there is always an opportunity for error*. What happens in the case of an unexpected event during the interaction? Failing gracefully and recovering has been studied in AI planning (Fox et al. 2006) where methods such as deciding when to replan may be applied, but these systems are often virtual and do not have physical robot concerns. If the robot is unable to act accordingly, then humans can be at risk from the robot's incorrect actions or by getting in the robot's way. Other accidents can take place including people falling down and recklessly running into things, especially when the interactions involve the elderly (Faria et al. 2015) and young children.

Due to the unexpected physical safety concerns, solutions for this challenge will need to integrate AI with the robot's hardware. An unexpected collision cannot simply be represented by a single state because the location of impact may affect the robot's stimuli and response. Robots such as U-Bot and BigDog are designed to maintain their balance at strong levels of impact while moving in the pushed direction; such recovery techniques have been identified by reinforcement learning (Kuindersma, Grupen, and Barto 2011). This both reduces the force against the human or object that hit the robot and avoids the danger of the robot falling over (onto other unsuspecting humans). Besides robots maintaining balance, other hardware that can assist with sensing unexpected contact are hybrid cartesian force/impedence controllers with energy tanks for passive contact with external forces (Schindlbeck and Haddadin 2015) and robotic skin, a full-chassis sensor that receives electric current from a point of contact like a touch-screen interface (Silvera-Tawil, Rye, and Velonaki 2015). If these hardware features become mainstream or standards, then it will be critical for the AI-HRI community to determine how to use their information.

Failure in interactions can also pose additional safety concerns when emotions and ethics are considered. If a robot is unable to assist the human properly, then her trust in it may decrease. Humans have been empirically shown to have different expectations of robots than fellow humans (Kwon, Jung, and Knepper 2016), but humans are willing to accept suggestions from robots if they contribute to better efficiency (Gombolay et al. 2014). Likewise, what happens if a human fails to do her part in an interaction or misleads the robot? Trust, like other morals, can be represented using higher-order logic in a knowledge base (Scheutz, Malle, and Briggs 2015). Can affecting these relationships between users and robots cause one to feel unsafe around the other? Lastly, in some cases of failure, there is no easy recovery option and an undesirable consequence will result from each one (Blass and Forbus 2015); such ethical questions are still complicated matters for humans in general.

5 Discussion

As the presence of robots increases in the real world and they have more direct interactions with humans outside of caged-off areas, it is important that the AI-HRI community expands its areas of study from improving the user experience to also considering its safety. We introduced three categories of challenges within the realm of safety in AI-HRI: the omission of humans from the robot's modeling, humans' lack of understanding of the robot's models and interior thoughts, and need to prepare for failures in execution with recovery. For each category, we showed how recent research can already be applied in these directions towards the proposed approaches. There is also an opportunity to consider literature in some areas of AI such as metareasoning (Cox and Raja 2011), which derives explanations for action decisions, and mixed-initiative planning (Bresina et al. 2005), which studies cooperative plan development between a human and machine. As it goes hand-in-hand with the quality of the user experience, improving the safety of HRI when AI has control is a present challenge worth considering.

Acknowledgments This work was supported by National Science Foundation Grant No. IIS-1405550

References

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in AI safety. *CoRR* abs/1606.06565.

Baraka, K.; Paiva, A.; and Veloso, M. 2015. Expressive lights for revealing mobile service robot state. In *Proc. of AAAI 2015 Fall Symposium on AI-HRI*, 17–23.

Blass, J. A., and Forbus, K. D. 2015. Moral decision-making by analogy: Generalizations versus exemplars. In *29th AAAI*, 501–507.

Bresina, J. L.; Jónsson, A. K.; Morris, P. H.; and Rajan, K. 2005. Mixed-initiative planning in MAPGEN: Capabilities and shortcomings. Technical report, NASA.

Cakmak, M., and Thomaz, A. L. 2012. Designing robot learners that ask good questions. In *7th ACM/IEEE HRI*, 17–24.

Cox, M. T., and Raja, A. 2011. *Metareasoning: Thinking About Thinking*. MIT Press.

Darling, K.; Nandy, P.; and Breazeal, C. 2015. Empathic concern and the effect of stories in human-robot interaction. In *24th IEEE Ro-Man*, 770–775.

Dragan, A., and Srinivasa, S. 2014. Integrating human observer inferences into robot motion planning. *Autonomous Robots* 37(4):351–368.

Enright, J., and Wurman, P. 2011. Optimization and coordinated autonomy in mobile fulfillment systems. In *AAAI Workshops: Automated Action Planning for Autonomous Mobile Robots*.

Erol, K.; Hendler, J.; and Nau, D. S. 1994. HTN planning: Complexity and expressivity. In *12th AAAI*, 1123–1128.

Faria, D. R.; Vieira, M.; Premebida, C.; and Nunes, U. 2015. Probabilistic human daily activity recognition towards robot-assisted living. In *24th RO-MAN*, 582–587.

Fox, M.; Gerevini, A.; Long, D.; and Serina, I. 2006. Plan stability: Replanning versus plan repair. In *16th ICAPS*, 212–221.

Freedman, R. G., and Fukunaga, A. 2015. Integration of planning with plan recognition using classical planners (extended abstract). In *Proc. of AAAI 2015 Fall Symposium on AI-HRI*, 48–50.

Freedman, R. G., and Zilberstein, S. 2016. Using metadata to automate interpretations of unsupervised learning-derived clusters. In *Proc. of IJCAI Workshop BeyondLabeler*.

Gombolay, M.; Gutierrez, R.; Sturla, G.; and Shah, J. 2014. Decision-making authority, team efficiency and human worker satisfaction in mixed human-robot teams. In *14th RSS*.

Hayes, B., and Scassellati, B. 2015. Effective robot teammate behaviors for supporting sequential manipulation tasks. In *IEEE/RSJ IROS*, 6374–6380.

Hirose, J.; Hirokawa, M.; and Suzuki, K. 2014. Robotic gaming companion to facilitate social interaction among children. In *23rd IEEE Ro-Man*, 63–68.

Hoffman, G., and Weinberg, G. 2010. Gesture-based human-robot jazz improvisation. In *IEEE ICRA*, 582–587.

Horstmann, M., and Zilberstein, S. 2003. Automated generation of understandable contingency plans. In *Proc. of ICAPS Workshop on Planning Under Uncertainty and Incomplete Information*.

Kim, B.; Rudin, C.; and Shah, J. A. 2014. The Bayesian case model: A generative approach for case-based reasoning and prototype classification. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *27th NIPS*. Curran Associates, Inc. 1952–1960.

Knepper, R. A.; Tellex, S.; Li, A.; Roy, N.; and Rus, D. 2015. Recovering from failure by asking for help. *Autonomous Robots* 39(3):347–362.

Knight, H., and Simmons, R. 2013. Estimating human interest and attention via gaze analysis. In *IEEE ICRA*, 4350–4355.

Kory, J., and Breazeal, C. 2014. Storytelling with robots: Learning companions for preschool children's language development. In *23rd IEEE Ro-Man*, 643–648.

Kuindersma, S.; Grupen, R. A.; and Barto, A. 2011. Learning dynamic arm motions for postural recovery. In *11th IEEE-RAS Humanoids*, 7–12.

Kwon, M.; Jung, M. F.; and Knepper, R. A. 2016. Human expectations of social robots. In *11th ACM/IEEE HRI*, 463–464.

Lasota, P. A.; Rossano, G. F.; and Shah, J. A. 2014. Toward safe close-proximity human-robot interaction with standard industrial robots. In *12th IEEE CASE*, 339–344.

Levine, S., and Williams, B. 2014. Concurrent plan recognition and execution for human-robot teams. In *24th ICAPS*, 490–498.

Maeda, G.; Ewerton, M.; Lioutikov, R.; Ben Amor, H.; Peters, J.; and Neumann, G. 2014. Learning interaction for collaborative tasks with probabilistic movement primitives. In *14th IEEE-RAS Humanoids*, 527–534.

Mainprice, J.; Rafi, H.; and Berenson, D. 2015. Predicting human reaching motion in collaborative tasks using inverse optimal control and iterative re-planning. In *IEEE ICRA*, 885–892.

McFarland, M. 2016. 300-pound mall robot runs over toddler. CNN Money.

Miller, D.; Sun, A.; Johns, M.; Ive, H.; Sirkin, D.; Aich, S.; and Ju, W. 2015. Distraction becomes engagement in automated driving. *Proc. of the Human Factors and Ergonomics Society Annual Meeting* 59(1):1676–1680.

Mirsky, R., and Gal, Y. K. 2016. SLIM: Semi-lazy inference mechanism for plan recognition. In *25th IJCAI*, 394–400.

Mörtl, A.; Lawitzky, M.; Kucukyilmaz, A.; Sezgin, T. M.; Basdogan, C.; and Hirche, S. 2012. The role of roles: Physical cooperation between humans and robots. *The International Journal of Robotics Research* 31(13):1656–1674.

Nguyen, A.; Yosinski, J.; and Clune, J. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE CVPR*, 427–436.

Nourbakhsh, I. R.; Sycara, K.; Koes, M.; Yong, M.; Lewis, M.; and Burion, S. 2005. Human-robot teaming for search and rescue. *IEEE Pervasive Computing* 4:72–78.

OSTP, and CMU. 2016. Workshop on safety and control for artificial intelligence. OSTP and CMU.

Pourmehr, S.; Thomas, J.; and Vaughan, R. 2016. What untrained people do when asked "make the robot come to you". In *11th ACM/IEEE HRI*, 495–496.

Price, R. 2016. Microsoft is deleting its AI chatbot's incredibly racist tweets. Business Insider.

Reben, A., and Paradiso, J. 2011. A mobile interactive robot for gathering structured social video. In *19th ACM Multimedia*, 917–920.

Rosenthal, S., and Veloso, M. 2012. Mobile robot planning to seek help with spatially-situated tasks. In *26th AAAI*, 2067–2073.

Scheutz, M., and Arnold, T. 2016. Are we ready for sex robots? In *11th ACM/IEEE HRI*, 351–358.

Scheutz, M.; Malle, B.; and Briggs, G. 2015. Towards morally sensitive action selection for autonomous social robots. In *24th IEEE Ro-Man*, 492–497.

Schindlbeck, C., and Haddadin, S. 2015. Unified passivity-based cartesian force/impedance control for rigid and flexible joint robots via task-energy tanks. In *IEEE ICRA*, 440–447.

Shafti, A.; Althoefer, K.; Orlandini, A.; Cesta, A.; Maurtua, I. n.; and Wurdemann, H. 2016. Workshop on planning, scheduling and dependability in safe human-robot interactions. Four by Three.

Silvera-Tawil, D.; Rye, D.; and Velonaki, M. 2015. Artificial skin and tactile sensing for socially interactive robots: A review. *Robotics and Autonomous Systems* 63, Part 3:230–243. Advances in Tactile Sensing and Touch-based Human Robot Interaction.

Sirkin, D.; Mok, B.; Yang, S.; and Ju, W. 2015. Mechanical ottoman: How robotic furniture offers and withdraws support. In *10th ACM/IEEE HRI*, 11–18.

Unhelkar, V. V.; Pérez-D'Arpino, C.; Stirling, L.; and Shah, J. A. 2015. Human-robot co-navigation using anticipatory indicators of human walking motion. In *IEEE ICRA*, 6183–6190.

Williams, T.; Briggs, P.; Pelz, N.; and Scheutz, M. 2014. Is robot telepathy acceptable? investigating effects of nonverbal robot-robot communication on human-robot interaction. In *23rd IEEE Ro-Man*, 886–91.

Yang, S.; Mok, B. K.-J.; Sirkin, D.; Ive, H. P.; Maheshwari, R.; Fischer, K.; and Ju, W. 2015. Experiences developing socially acceptable interactions for a robotic trash barrel. In *24th IEEE Ro-Man*, 277–284.

Zilberstein, S. 2015. Building strong semi-autonomous systems. In *Proc. of 29th AAAI*, 4088–4092.