Optimizing for Blame Allocation in Human-Robot Systems

Wendy Xu
Austin Whitesell
William D. Smart
Robotics Program
Oregon State University
Corvallis, OR 97331, USA
xuwe@oregonstate.edu
whitesea@oregonstate.edu
smartw@oregonstate.edu

Abstract
In this paper, we discuss how robots and humans can work together in the treatment of highly infectious diseases such as Ebola, and pose questions about how blame will be allocated among the human and robot team members in the case of a failure. We discuss how blame might be estimated in this setting, and how we might optimize the allocation of sub-tasks to human and robot team members based on this estimation, rather than on traditional task-based metrics such as expected ask completion time.

Author Keywords
Human-Robot Interaction; Task Allocation; Blame H.5.3 Group and Organization Interfaces

ACM Classification Keywords
H.5.3 [Group and Organization Interfaces]: Computer-supported cooperative work

Introduction
We have recently begun a project that looks at the use of robotics and automation to improve patient outcomes in a highly-infectious disease treatment facility. Using the 2014–16 outbreak of Ebola Virus Disease in West Africa as a model, we are experimenting with how we can integrate robots and other automation into the existing protocols and treatment procedures used by Non-Governmental Organi-

zations like Mèdecins Sans Frontiéres. In these settings, robots can be used to both protect the health care workers from infection by performing some of the more dangerous tasks, and also to free them from mundane tasks so that the can concentrate more on patient care.

However, our initial work in this area has highlighted that many of the issues that we will face are not technical ones [5]. We anticipate social issues will be felt more acutely because of the high-stress nature of the environment, and that managing this effectively will be more important than any of the component technologies involved. Particularly, we examine how blame is assigned and how blame may relate to acceptance, trust, and accountability, all factors integral to a strong collaboration. While we are initially focused on Ebola and on under-resourced environments, we strongly believe that the lessons we learn here will transfer to other collaborations, in more resourced environments, and settings other than the treatment of highly-infectious diseases.

In this paper, we intentionally raise more questions than we answer. We begin with some background on the recent Ebola outbreak and our project. We then go on to identify some of the questions that we are currently thinking about, and discuss our plans for addressing them. Finally, we conclude with a brief discussion of how the work we plan can generalize to other settings.

Background

Ebola Virus Disease (EVD) is an often fatal hemorrhagic fever, transmitted through direct contact with infected bodily fluids. Initial symptoms resemble those of influenza, with fatigue, fever, sore throat, and muscular pain being common. These are typically followed with internal and external bleeding, vomiting, and diarrhea. Mortality rates are typically around 50%, but can vary widely depending on the quality and quantity of medical care available. Treatment of EVD is typically supportive: oral and intravenous rehydration, pain and nausea management, and anxiety minimization [8].

Since EVD can only be transmitted through direct contact with infected bodily fluids, most protective measures focus on preventing this contact. Ebola treatment units (ETUs), such as those set up by Mèdecins Sans Froniéres (MSF) [6] rely on strict separation between ares with and without virus contamination. Health care workers wear extensive personal protective equipment (PPE) comprising face masks, gowns, medical scrubs, boots, and several layers of surgical gloves. An complete set of PPE clothing in an MSF facility costs approximately $60 [1]. It is used once, and then discarded.

The Recent Ebola Outbreak in West Africa

The largest EVD outbreak to date started in Guinea, West Africa, in December of 2013. In 2014 and 2015 it spread to the neighboring countries of Liberia and Sierra Leone, with minor outbreaks elsewhere. As of January 2016, the outbreak was considered to be under control, but not yet over [9]. The outbreak had over 15,000 confirmed cases and over 11,000 deaths, a mortality rate of approximately 74%. In the three most-affected countries (Guinea, Liberia, and Sierra Leone), there were and additional 13,386 suspected and probable cases. Although this could bring the mortality rate as low as 40%, this is still a staggering number. In the four Western countries (Spain, United Kingdom, Italy, and United States) in which cases were reported, there were seven confirmed cases, and only a single fatality.

Conditions in the affected West African countries made it difficult for health care workers to effectively treat victims, while still safeguarding their own health. Temperatures in

the affected areas were often close to 100° Fahrenheit with extremely high humidity. This restricted health care workers wearing PPE to work shifts as short as 45 minutes[1], to avoid heat-related illnesses. This meant that many routine tasks, such as patient monitoring and replacement IV fluid containers, occupy a full shift. When the time to don and doff the PPE is often greater then the time spent in it, patient care was severely compromised. The limited number of trained health care staff and the necessarily short shift times combined to severely restrict the amount of time a health care worker could spend with a given patient [7].

Our Approach and Experimental Setup

Our approach to using robots and automation in this setting is quite straightforward: we will use robots to reduce the exposure to EVD of health care workers, performing mundane tasks that will allow them to focus on the patients. This will allow them to make the best use of the time that they have in their shift and, hopefully, improve patient outcomes. We are explicitly not looking at using robots to directly interact with patients, and restrict ourselves to considering other tasks that humans might do while in PPE, such as fetching and carrying supplies, cleaning, stand-off monitoring, and the like.

Our test environment will be modeled after the ETUs used by MSF in the recent outbreak. We will build the treatment facility and staff it with student workers and nursing students from the local community college. We'll also replicate care guidelines by breaking them down into tasks and subtasks, and the dependencies between them (task A must be complete before task B). We can take this representation and attempt to optimize the workflow, assigning tasks and sub-tasks to humans in an attempt to minimize the amount of time that anyone is standing idle. Once this is done, we will introduce a robot (or robots) to the system to automate appropriate sub-tasks. These studies will be done with a PR2 robot or a similar mobile robot capable of interacting with the environment. While the PR2 may not be ideal for an application such as this, we believe it will adequately show the effectiveness of automating sub-tasks. This allows us to have more agents performing work, increasing the amount that can be done, and to allow the humans to preferentially focus on patient-centered tasks. A key feature of our approach will be to do this task allocation *dynamically*, responding as tasks are completed, rather than determining a fixed allocation ahead of time. This will allow us to be more responsive to tasks that can take a varying length of time more effectively, or robots that are starting to perform poorly because of some failure.

Potential Challenges

While the technical challenges of getting robots to work effectively in these settings should not be underestimated, we believe that effectively integrating them into the existing workflow of human health care workers will be a harder problem. We are taking a new technology, and trying to insert it into a high-stress, life-or-death working environment. Health care workers are over-worked and at a real risk of contracting Ebola. Patients are often terrified, in unfamiliar surroundings, and in great physical discomfort. The reaction to any mistakes made by the robot will be amplified, both for the patient and the health care workers. Since the health care workers will be performing skilled work, there is evidence to suggest that they will preferentially blame the robot when something goes wrong [3]. We are taking a stressful situation, and making it potentially worse by introducing a new, scary piece of technology that the humans have never seen before. This makes it critical for us to understand what the effects of this technology will be, with respect to the existing inter-personal dynamics at play in the Ebola treatment setting.

The Research Questions

In this paper, we will focus on two questions: (1) who (or what) gets blamed when something goes wrong in this setting; and (2) how can we use this understanding of the blame game to allocate tasks such that blame will be assigned to the ideal agent when something does go wrong? Across all of these questions, we are interested in thinking about the patients, the health care workers, and the robots as members of a single system. This, we claim, makes our setting different (and, perhaps, more interesting) than other in the literature, since some of the humans in the system (the patients) have no direct control of the robots, and often little control over anything else in the environment.

Who's to Blame?

The interactions that we are interested in comprise three types of actors: patients, health care workers, and robots. If we make the assumption that the patient will never be held responsible for the failure to accomplish a task, that leaves us with two actors to blame. Similarly, if we assume that the robot will not assign blame on a task failure, then we have two actors that can apportion blame if something goes wrong. The question, then, is who blames whom, and under what circumstances. It seems clear that if one of the actors is solely responsible for a task, the it should be held to blame if that task goes wrong. However, our situation is somewhat muddied, since we assume that the robot may be under the control of a human who is not physically present in the room. In the case where the robot causes the failure, is it the robot, the operator, or some combination of the two that are assigned blame? Kaniarasu and Steinfeld showed that, when a human-robot team is working on a skill-based task, something that the human practices and perfects, then there is a tendency to blame the robot for any problems that occur [4]. In a similar vein, Groom et al. ran an experiment where it either blamed itself, the whole team, or the human it was working with when a human-robot collaboratve task failed. They found that people preferred it when the robot blamed itself instead of the team, and the team instead of the human [3]. Is it a general tendency to blame the actor that we have less of an emotional relationship with? What if the medical personnel consistently work with the robot and are now familiar, will they blame the new person that joins the team?

Who's Ideally to Blame?

Regardless of which actors will be to blame, what are the positive and negative consequences of that blame assignment? Kaniarasu and Steinfeld found that blaming the robot led to a decreased trust in the robot system. This might be problematic for us, since there is evidence the suggest that people are less willing to let an automated system complete a task for them if they do not trust it [2]. However, since we are considering environments where people are not likely to have seen or worked with a robot before, how will this affect their level of trust and sway how they assign blame? Will they ever trust the robot, even if it performs perfectly? Are they starting from a place of such profound mistrust that it is impossible for them to ever get beyond that state?

If we choose to ideally allocate sub-tasks to the robot when their failure would not be seen as significant or even notices, there maybe an obvious tension; we want to allocate sub-tasks in a way that will increase the efficiency of the human health care worker but, at the same time, reduce the chances that a failure is blamed on the robot. These two things might be directly at odds with each other.

Another drawback would be blaming robots when teams or humans actually want to be responsible for and credited for the failures. For example, a nurse might want credit for a failure that accidentally leads to better patient care.

From another perspective, should we have the robot perform a risky task and absorb the blame when it goes wrong, rather than having the patient blame the health care worker and, consequently, lose trust in them? This is a similar argument, in spirit at least, to the one made by Elish that one of the roles of the human in any human-robot system is to act as a "moral crumple zone" to absorb any legal repercussions of the system. In our example, the situation is reversed and the robot would assume the blame resulting from the failure, rather than the human.

Allocating Tasks to Direct Blame?
Our approach to adding robots to the Ebola treatment setting is to model the current workflow, to automate some of the tasks, and then to optimize the task allocation in an attempt to improve patient outcomes. This general framework is not particularly novel, since especially when one considers traditional metrics over which to optimize, such as time to complete a given set of tasks. However, given the different implications of blaming the robot vs humans for task failures and regardless of what is the ideal blame taker, perhaps there is another optimization that we can perform. Instead of optimizing over time, or some similar metric, we optimize the task allocation to, let's say, *minimize the loss of trust in the robot system*, in addition to maximizing the improvement in measurable patient outcomes.

Before we can even get to the question of allocation and optimization, however, we must first figure out how to estimate blame. How do we consistently measure and assign blame when a sub-task fails? The assignment of blame is likely to be a highly subjective thing, and dependent on many factors outside of our control. Also, it is not likely to stay constant over time, unlike subjective measures like the time taken to perform a task. Suppose we attempt to estimate the likelihood of assigning blame to the robot on a failure by having a human observe it as it fails, and then asking them to assign blame. The more often they see a given robot fail at a task, the more likely they will be to assign blame to it. However, bringing in a fresh observer for each failure is not tenable.

Once we have a metric to measure how likely it is for a robot to be blamed for a failure, we can combine that with the probability of the failure itself to come up with an estimate of how risky it would be to allocate that sub-task to the robot, and use that in our optimization framework.

Conclusion

As promised, in this paper we raised a number of questions about the allocation of blame in a human-robot system without really answering any of them. Our goal with this paper is simply to raise some of these questions, and spur debate. We believe that our setting, the treatment of highly infectious diseases like Ebola, exacerbates the problem of blame because of it's inherent high-risk nature. Failure and contingencies in an environment containing Ebola virus is a serious thing and this will, we believe, cause blame to be attached more strongly than in other, more benign, environments.

We also introduced the idea of allocating tasks based on the likelihood of blame being assigned, rather than other more traditional metrics, such as estimated time to completion. Although blame will be a hard thing to measure objectively, such a system could be used to ideally direct where blame should be assigned to try to preserve human trust in a robot system or to absorb the blame for a high-risk action, allowing the human health care worker to retain the trust of the patient under certain conditions.

Estimating the likelihood of blame will be hard, but we suspect that it will be even harder for settings like the one de-

scribed in this paper. An active Ebola outbreak is a terrifying thing, and everyone involved is under a great deal of stress and danger. This is sure to change the ways in which people assess risk and assign blame for (perceived) failures. While we can develop techniques to assess and measure the ascription of blame in a safe laboratory setting, it remains an open question how well these techniques will carry across to an actual infectious disease outbreak, where real lives are on the line.

Acknowledgements

The work described in the paper is supported in part by the US National Science Foundation, under awards IIS 1518652 and IIS 1540483, and by the US National Institute of Biomedical Imaging and Bioengineering, under award R01 EB024330-01. The authors would also like to thank Armand Sprecher of Médecins Sans Frontières for his help and support.

REFERENCES

1. 2014. Symposium on Advancement of Field-Robots for Ebola Response (SAFER). Worcester, MA. This workshop was conducted under Chatham House Rules.
2. Munjal Desai, Kristen Stubbs, Aaron M. Steinfeld, and Holly Yanco. 2009. Creating Trustworthy Robots: Lessons and Inspirations from Automated Systems. In *Proceedings of the International Symposium on New Frontiers in Human-Robot Interaction*. Edinburgh, Scotland.
3. Victoria Groom, Jimmy Chen, Theresa Johnson, F. Arda Kara, and Clifford Nass. 2010. Critic, Compatriot, or Chump?: Responses to Robot Blame Attribution. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Osaka, Japan, 211–218.
4. Poornima Kaniarasu and Aaron M. Steinfeld. 2014. Effects of Blame on Trust in Human Robot Interaction. In *The IEEE International Symposium on Robot and Human Interactive Communication (Ro-Man)*. 850–855.
5. Kory Kraft and William D. Smart. 2016. Seeing is Comforting: Effects of Teleoperator Visibility in Robot-Mediated Health Care. In *Proceedings of the IEEE/ACM International Conference on Human Robot Interaction (HRI)*. Christchurch, New Zealand.
6. Mèdecins Sans Frontiéres. 2014. Interactive: Explore an Ebola Care Center. (2014). `http://www.msf.org/article/interactive-explore-ebola-care-centre`.
7. World Health Organization. 2015. Factors that Contributed to Undetected Spread of the Ebola Virus and Impeded Rapid Containment. (Jan. 2015). `http://www.who.int/csr/disease/ebola/one-year-report/factors/en/`.
8. World Health Organization. 2016a. Ebola Virus Disease. (Jan. 2016). `http://www.who.int/mediacentre/factsheets/fs103/en/`.
9. World Health Organization. 2016b. WHO Director-General addresses the Executive Board. (Jan. 2016). `http://www.who.int/dg/speeches/2016/executive-board-138/en/`.