Nuancing the human-robot relation: Intelligent disobedience and human failures in robot teleoperation

Kavyaa Somasundaram
kavyaa.somasundaram@oru.se
Orebro university
Sweden

Katherine Harrison
katherine.harrison@liu.se
Linköping University
Sweden

Andrey Kiselev
andrey.kiselev@oru.se
Orebro university
Sweden

Amy Loutfi
amy.loutfi@oru.se
Orebro university
Sweden

ABSTRACT

Human-robot interaction in any form and application domains are susceptible to failures. Failures are of varied definitions and types. The severity of the failure is high for teleoperated tasks. In this paper, we focus on the less explored type of failure called human induced interaction failure using the concept of Intelligent Disobedience (ID). ID was first used as a method for training guide dogs, where the dogs disobey the user's commands that are dangerous to perform. Imparting this behaviour of disobedience to manage human induced interaction failures consists of numerous social, cultural and ethical aspects to consider. This paper discusses a novel framework based on ID to manage human induced interaction failures. Also, we discuss a nuanced approach involved in robot disobedience considering the different social and cultural aspects.

CCS CONCEPTS

• **Computer systems organization → Robotics**; • **Human-centered computing → Interaction design theory, concepts and paradigms**.

KEYWORDS

intelligent disobedience, failures in HRI, interaction failures, human errors

1 INTRODUCTION

The field of study known as human-robot interaction (HRI) involves creating robots that can interact with humans in a way that is natural and intuitive.[1]. Teleoperated robot platforms with a certain degree of autonomy are increasing both in social and industrial application domains where human-robot co-existence is necessary. Trust and safety are essential aspects of building a reliable relationship within a human-robot team [12]. An important factor that influences the effectiveness of this relationship is to manage and prevent failures which are inevitable. These failures can occur due to a variety of reasons, including design and technical limitations of the robot [10], misunderstandings in language and social cues [7], and mismatches in the abilities and expectations of the human and the robot [14]. Thus, the need to understand and design better failure management systems is significant to create an intuitive human-robot interaction.

The definition of failure that is followed in this paper is from Brooks, where he defines failure as " a degraded state of the ability which causes the behaviour or service being performed by the system to deviate from the ideal, normal or correct functionality" [2]. The classification of failures that is opted in this paper is from [10], where the broad classification of failures is of two types namely, technical failures and interaction failures. Interaction failures being the less explored are further subdivided into failures caused by social norm violations, human errors and environmental or other agents. In this work, we focus on the interaction failures that are induced by human errors such as mistakes, slips, lapses and deliberate violations. Here, mistakes are conscious errors such as choosing the wrong method for a task, slips are unconscious errors such as choosing the wrong method unintentionally when attempting to do the right thing. Lapses refer to the attention and memory loss a human could have during a task and finally deliberate violations are actions that could jeopardize the task intentionally.

Intimating the user through communicating mediums such as vocal signals [3] or visual indicators in case of an error [13] is the most commonly used approach for managing interaction failures. These methods are not effective in preventing the failures and to the best of our knowledge, there is no available framework that deals with human induced interaction failures. Thus, this paper focuses on the above-mentioned research gap.

2 INTELLIGENT DISOBEDIENCE

Intelligent disobedience (ID) is a recent development in human-robot interaction (HRI). It pertains to a robot's ability to override a direct instruction from a human if it conflicts with ethical, moral, or safety considerations[4]. The ultimate goal is to ensure that robots are capable of taking actions that protect human well-being, even if it means disobeying a command. This concept is particularly relevant in scenarios where robots are employed in dangerous or high-stakes environments, such as in robot teleoperation. The idea behind ID originated from the training of guide dogs. Guide dogs are taught to disregard commands given by their handlers if those

ImpRR23 Workshop, HRI'23, March 13–16, 2023, Stockholm, Sweden

ACM ISBN 978-X-XXXX-XXXX-X/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

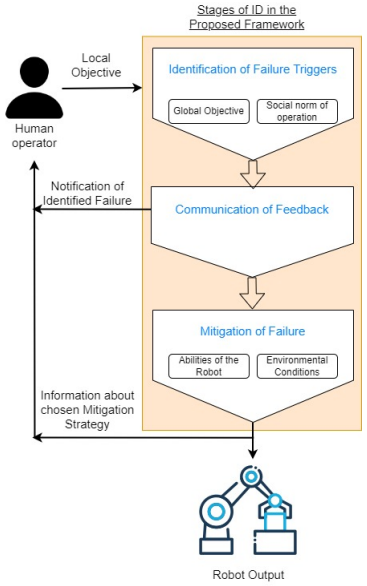

Figure 1: Stages in the Proposed Framework of ID for Managing Human Induced Interaction Failures

commands seem to place their handlers in danger. In this paper, we propose extending the concept of ID into a framework that prevents and mitigates interaction failures that result from human actions. To demonstrate how this concept would work in a teleoperating scenario, we present two hypothetical scenarios within the context of underwater robot teleoperation.

Scenario 1: During an underwater oil pipe inspection, the user located in a safe remote environment accidentally moves the robotic arms in the direction of the pipes at greater speed. The robot arm hits the pipe and as a result oil leakage occurred.

Scenario 2: While collecting samples in a swimming pool, the user looking at the monitor drives the robot forward. Suddenly, a human who is unaware of this situation takes a dive into the restricted area. Due to the general time delay with video monitoring, the operator is unaware of the situation and proceeds to drive the robot in the same direction. As a result, the robot collides with the human who took a dive.

The above two scenarios would be preventable if the robot was not designed to simply follow the instructions given by the human operator. By utilizing the concept of ID, these dangerous situations can be prevented.

3 PROPOSED FRAMEWORK TO PREVENT HUMAN INDUCED INTERACTION FAILURES

Intelligent Disobedience (ID) have been previously used in designing seeing eye robot as a replacement for guide dogs [11] and in AI rebel agents [5]. This framework takes inspiration from the above-mentioned studies and reverse-engineered the behaviour pattern of the guide dog that is trained to be intelligently disobedient. The framework consists of three stages which are shown in Figure 1.

Stage 1 - Identification of failure trigger: The control sequences initiated by the human operator are considered local objectives. Global objectives (e.g., reaching the destination location, not damaging the robot etc.,) and General norms of operation (e.g., always reducing the speed while turning/approaching an object etc.,) are given to the system as predefined knowledge. When there is a contradiction between the local objective and the predefined knowledge, the system identifies that the control sequence initiated by the human is a trigger to a failure event.

Stage 2 - Communication of feedback: When the human operator is unable to reason the behaviours exhibited by the robot, the effectiveness of human-robot collaborations takes a negative effect. Good feedback communication is essential in avoiding this issue. In the proposed system, the identified failure trigger and the choice of mitigation strategy that is selected at the later stages are communicated back to the user.

Stage 3 - Mitigation of failure: The last step in managing the human induced interaction failure is choosing what tools can be used to mitigate the identified failure trigger. Here, the abilities of the robot (e.g., the direction of motion possible by the robot, sensors available etc.,) are predefined knowledge of the system. Thus, the system makes use of the environmental conditions and the predefined knowledge to choose the appropriate mitigation strategy that is feasible.

If the proposed framework using ID is applied to the scenarios mentioned in Section 2, the outcomes would differ. In scenario 1, the robot would not move its arms closely towards the oil pipes at a higher speed and delegated the failure through inaction. In scenario 2, the system would have identified that a human is present in the scene and changed its direction of motion instead of bombarding directly with the human.

4 NUANCING THE HUMAN-ROBOT RELATION

Intelligent disobedience on the part of the robot potentially offers a way to temporarily override the human control of the interaction to avoid dangerous or unsafe situations. These situations are ones caused by human failures, which then cause – what at least appears to be – robotic failures (when the robot disobeys its instructions, although doing so for good reason). This section highlights some potential challenges in developing intelligent disobedience by discussing context-specificity and the interaction space. Despite some apparent similarities (underwater, teleoperated robots with human handlers), the two scenarios described in Section Two show important differences. The first involves only the operator and the robot, while the second involves another human who is accidentally involved and unaware of the test being performed. The first is in open water while the second is in a swimming pool suggesting significant differences in water clarity, proximity of the operator, and the likelihood of other bodies. The first would lead to material damage of a pipe, while the second to an injury of a human. In the first, the robot must act due to a mistake in using the controls while the second occurs through time delay which results in the operator being unaware of the other human. As such, they suggest that the successful use of intelligent disobedience is highly context specific.

Also relevant here are discussions on failures in HRI that have drawn attention to the different ways in which "failure" may be understood [9]. The scenarios described earlier in the paper define "failure" as firstly, causing damage to material objects (an oil pipe), and secondly, as injuring a human being. Both scenarios stop at the point of immediate mitigation of the task but do not consider the next step in the interaction. Intelligent disobedience may well allow a robot to prevent damage or danger, but it also challenges the relationship with its human counterpart. Teleoperated robot work requires a high level of trust and experience in working together with a robot. Disobedience as described here may impact that trust relation and the master-slave dynamic that currently shapes many interactions between humans and robots. As other studies have shown, human responses to failure mitigation depend heavily on pre-existing experience with robots, as well as proximity to the interaction (participant or bystander) [6], [8]. It is therefore possible that, in preventing a human failure, the robot provokes a failure in its relation to its human counterpart. In outlining this, we aim to widen the understanding of "failure" under consideration, by moving from the narrow time and space frame of the proscribed interaction to an ongoing relationship between humans and robots more akin to our everyday relations with familiar technologies.

It is fundamentally troubling to humans to have a robot "know better". The power balance between humans and robots is represented and explored in many popular forms, perhaps most famously in Asimov's "Three Laws of Robotics". Thus, a robot that knows better and acts upon that knowledge represents a potentially troubling agent in our world. It poses questions about how and when a robot should disobey, and demands sensitivity to the expertise and vulnerability of surrounding humans (both handler and bystander). Intelligent disobedience thus highlights the need for a more nuanced understanding of human-robot relations and an accompanying ethical framework to guide such research.

5 CONCLUSION

In human-robot interactions, both humans and robots will fail. To date, research has focused primarily on managing robot failures and how to mitigate these so that the human continues to interact. Less well understood are the effects of human failure on the interaction, especially in those cases where human failures result in dangerous or unsafe situations. In such cases, intelligent disobedience on the part of the robot could prevent danger or damage from occurring. The proposed framework increases reliability and explainability in a system which in turn will positively affect the effectiveness of the human-robot team. The act of robots disobeying human's instructions is a debate and we are aware of the social, cultural and ethical aspects involved. Can we rely on autonomous systems for making decisions, who would take responsibility for the system's action, will overriding the human's instructions have a negative impact on human's trust in the robot? These are interesting questions which do not have easy answers. Having open discussions and gaining different perspectives on this topic helps us in improving the system's design. Refining the proposed framework in terms of identifying the exact methods and tools that are required to implement each stage, by considering the mentioned social and cultural factors will be our future work. Intelligent disobedience poses important questions to the HRI community about the balance of power and the ongoing negotiations for trust between humans and robots.

6 ACKNOWLEDGEMENT

This work has been supported by the Industrial Graduate School Collaborative AI and Robotics funded by the Swedish Knowledge Foundation DNR:20190128. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation. DNR MMW 2019.0151.

REFERENCES

[1] C. Breazeal. 2004. Social interactions in HRI: the robot view. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 34, 2 (2004), 181–186. https://doi.org/10.1109/TSMCC.2004.826268

[2] Daniel J Brooks. 2017. HUMAN CENTRIC APPROACH TO AUTONOMOUS ROBOT FAILURES. *Ph.D. dissertation, Department of Computer Science, University. Lowell, MA* (2017), 24.

[3] Daniel J. Brooks, Momotaz Begum, and Holly A. Yanco. 2016. Analysis of reactions towards failures and recovery strategies for autonomous robots. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 487–492. https://doi.org/10.1109/ROMAN.2016.7745162 ISSN: 1944-9437.

[4] Ira Chaleff. 2015. *Intelligent disobedience: doing right when what you're told to do is wrong* (first edition ed.). Berrett-Koehler Publishers, Inc, Oakland, CA.

[5] Alexandra Coman and David W Aha. 2018. AI rebel agents. *AI magazine* (2018), 16–26. https://doi.org/10.1609/aimag.v39i3.2762

[6] Filipa Correia, Carla Guerra, Samuel Mascarenhas, Francisco S. Melo, and Ana Paiva. 2018. Exploring the Impact of Fault Justification in Human-Robot Trust. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (Stockholm, Sweden) *(AAMAS '18)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 507–513.

[7] Petra Gieselmann and Mari Ostendorf. 2007. Problem-Sensitive Response Generation in Human-Robot Dialogs. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*. Association for Computational Linguistics, Antwerp, Belgium, 219–222. https://aclanthology.org/2007.sigdial-1.38

[8] Haley N. Green, Md Mofijul Islam, Shahira Ali, and Tariq Iqbal. 2022. Who's Laughing NAO? Examining Perceptions of Failure in a Humorous Robot Partner. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 313–322. https://doi.org/10.1109/HRI53351.2022.9889353

[9] Katherine Harrison, Kavyaa Somasundaram, and Amy Loutfi. 2023. *The Imperfectly Relatable Robot: an interdisciplinary approach to failures in human-robot relations, In What that Robot Made me Feel* (forthcoming ed.). MIT Press.

[10] Shanee Honig and Tal Oron-Gilad. 2018. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in Psychology* 9, JUN (2018). https://doi.org/10.3389/fpsyg.2018.00861

[11] Reuth Mirsky and Peter Stone. 2021. The seeing-eye robot grand challenge: Rethinking automated care. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS* 1, Aamas (2021), 28–33. ISBN: 9781713832621.

[12] Aaron Pereira, Mareike Baumann, Jonas Gerstner, and Matthias Althoff. 2022. Improving Efficiency of Human-Robot Coexistence While Guaranteeing Safety: Theory and User Study. *IEEE Transactions on Automation Science and Engineering* (2022), 1–14. https://doi.org/10.1109/TASE.2022.3201970 Conference Name: IEEE Transactions on Automation Science and Engineering.

[13] Satragni Sarkar, Dejanira Araiza Illan, and Kerstin Eder. 2018. Effects of Faults, Experience, and Personality on Trust in a Robot Co-Worker. (2018). arXiv:1703.02335

[14] Leimin Tian, Pamela Carreno-Medrano, Aimee Allen, Shanti Sumartojo, Michael Mintrom, Enrique Coronado Zuniga, Gentiane Venture, Elizabeth Croft, and Dana Kulic. 2021. Redesigning Human-Robot Interaction in Response to Robot Failures: A Participatory Design Methodology. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 57, 8 pages. https://doi.org/10.1145/3411763.3443440