Intelligent Disobedience: A Novel Approach for Preventing Human Induced Interaction Failures in Robot Teleoperation

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

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

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

ABSTRACT

Failures are natural and unavoidable events in any form of interaction, especially in human-robot interactions (HRI). Throughout the literature, the definition and classification of failures are diverse, depending on the source and application domain. However, the tolerance to the aftereffect of these failures is low in teleoperation due to its unstructured application domains. One such type of failure is called human induced interaction failure. This is an interesting and often overlooked failure type, due to the perspective that robots are designed always to obey the instructions given by the human operators. Regardless of the degree of automation that the robot is equipped with. But what if the instructions provided are faulty, dangerous, or misleading. This paper addresses the above mentioned research gap. It introduces a framework based on the concept of Intelligent Disobedience (ID), derived from guide dog training methods, to manage human induced interaction failures in teleoperation scenarios.

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

ACM Reference Format:
Kavyaa Somasundaram, Andrey Kiselev, and Amy Loutfi. 2023. Intelligent Disobedience: A Novel Approach for Preventing Human Induced Interaction Failures in Robot Teleoperation. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23 Companion), March 13–16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3568294.3580060

1 INTRODUCTION

Human-robot co-existence in everyday space is steadily increasing in robot teleoperation, for robots that are teleoperated in professional as well as personal contexts [15]. An emerging characteristic of these teleoperated robot platforms is equipping the robot with semi-autonomous features such as grasping an object [8], navigating to a specific way point [14], where certain tasks can be performed autonomously by the robot even in teleoperated mode. Like in all human-robot interaction scenarios, failures are a natural and undeniable part of the interaction. Teleoperated HRI contexts are no exception.

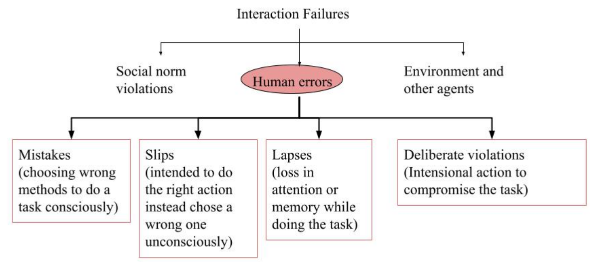

Figure 1: Classification of Interaction Failures

Throughout the literature, there are several definitions of failure; in this paper, we adopt the definition given by Brooks as "a degraded state of ability which causes the behavior or service being performed by the system to deviate from the ideal, normal or correct functionality" [3]. Failures are often denoted by other terms such as errors and faults, but faults lead to errors which then, in turn, leads to a failure [10]. A good example to understand the chain of events leading to a failure is a tool-picking robot failing to pick up a tool/object (failure) as a result of an error in the perception system (error) due to poor illumination (fault).

Failures in HRI are broadly classified based on the type (technical, physical, etc.,) [9], severity [11], recoverability [17], and interaction errors [5]. Yet, dealing with failures is largely centered around the prediction and mitigation of physical failures (effectors, sensors, software, and hardware). In this paper, we focus on a branch of failure that is less explored, termed interaction failures, by adopting the classification paradigm proposed in [10]. In their paper, the authors classify interaction failures into three types: social norm violations (behaviors that neglect common rules of operation), sudden changes in the environment or other agents, and human errors such as mistakes (conscious errors), slips (unconscious errors), memory lapses (loss of attention or memory), deliberate violations (wrongful intention) as shown in Figure 1.

Unlike physical failures, interaction failures are challenging to detect directly as they will not display external symptoms (e.g.: loose attachment of wheels in a robot) or alarms (e.g.: warning lights for sensor malfunction). The existing approaches to managing failures in HRI focus on communicating failures by designing visual indicators such as displaying emojis with an explanation when an error occurs [18] or by using audio and speech signals

HRI '23 Companion, March 13–16, 2023, Stockholm, Sweden

ACM ISBN 978-1-4503-9970-8/23/03.
https://doi.org/10.1145/3568294.3580060

to grab the user's attention [4]. While, approaches for mitigating failures include a complete halt of action, asking for help from the human handler, recovering by itself, etc. ([20], [2], [16]). The above-mentioned methods are modular, do not provide solutions for preventing interaction failures, and do not account for methods to address human induced interaction failures.

In this paper, we propose a holistic approach to manage human induced interaction failures while teleoperating a robot by adopting the concept of Intelligent Disobedience (ID) and thus propose a novel architecture defining the stages involved in implementing ID.

2 INTELLIGENT DISOBEDIENCE

Culturally, disobedience is an act to be punished, and obedience is an act that should be rewarded. This perspective is widely reflected in designing robotic behaviour, where most of the human–robot collaboration is designed to be a master–slave relationship. The aspect that this type of dynamics fails to account for is that humans tend to commit mistakes. In a teleoperation context, these mistakes might lead to disastrous outcomes depending on the application.

The concept of Intelligent Disobedience is derived from the practice of training a guide dog. Guide dogs are used to help visually impaired people navigate and perform daily tasks with ease. They are extremely obedient to their handlers and are aware of the general norms of the real world. One of such essential training given to these guide dogs is for them to behave intelligently disobedient [6]. which can be explained with an example. Let us assume that the handler proceeds to cross the road and that the signal is green for pedestrians. In this scenario, there might be multiple parallel crossing roads and other bystanders involved in the situation. Now, if a cyclist appears to come in the opposite direction towards the handler with increased speed, aiming to cross the road as well, the dog must disobey and refuse to move forward even if the handler insists that the dog do so. Instead, the dog would navigate the handler outside of the direction of the cyclist, while avoiding bumping into other bystanders in the environment. This kind of behaviour is termed as Intelligent Disobedience (doing right when what you are instructed is wrong/dangerous).

This way of intelligently making informed decisions would be a technical skill that could be transferred to a robot to create an effective coordinated relationship and avoid human induced interaction failures during teleoperation.

The concept of ID is discussed mainly in the creation of assistive robots, which is an attempt to replace the service dog with a robot and was first proposed by Tachi and Komoriya in the context of obstacle avoidance [19]. Pilot studies with NaO robots in the context of robot reasoning have been shown in this study [1], where the NaO robot is asked to jump a table and it refuses to do so. However, the same concept cannot be extended to other scenarios, as the robot lacks explicit knowledge to make informed judgments. Another area where ID has recently been applied is with handheld robots that use gaze detection for intention recognition [12]. First, the author explored the idea of using rebellion and disobedience as useful tools to correct user actions. Furthermore, the paper investigates the effectiveness of collaboration and examines how users react when the robot corrects their actions. The author concluded that participants rated feeling frustrated when the robot rebelled

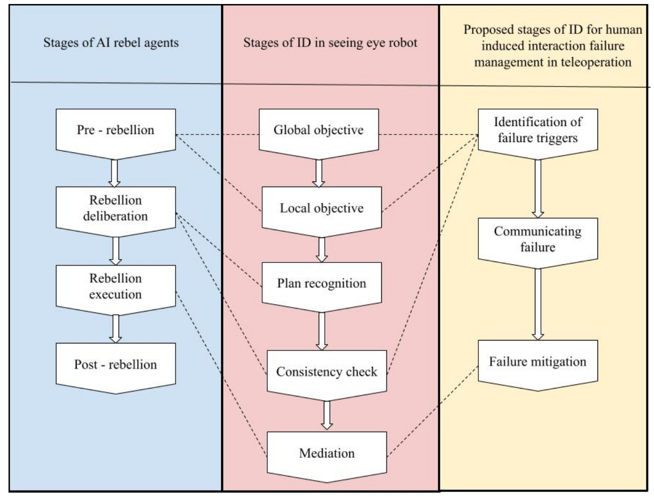

Figure 2: Illustrates the different stages of the proposed framework and the dotted lines represent the similarities between each stage defined in the proposed framework in contrast with the other two models

and nudged the user towards the correct action. These studies have partially implemented ID with task knowledge alone and lack generalization to other scenarios within the same task. To the best of our knowledge, there is no literature available on the application of ID in the teleoperation domain for managing human induced interaction failures.

3 PROPOSED ARCHITECTURE

Implementing ID for a teleoperating task needs to address two important questions, when is the right time to disobey, and what is the right task to do after disobeying. The inspiration for our framework was drawn from two works, first, [13] who have proposed a framework for ID in the context of the seeing eye robot. The author has defined the different stages of ID as the global objectives (overall goals), local objectives (current goal of the handler), followed by plan recognition (understanding of the method intended to do a task), consistency check (checking if there is a conflict or not), and finally mediation (making decisions about the next steps of mitigation).

The other study is by [7] where the ID architecture is proposed in the context of AI rebellion agents. The behavior of agents is divided into prerebellion (changes in the environment and other agents are monitored), rebellion deliberation (set of conditions to decide whether to trigger rebellion behavior or not), rebellion execution, and finally post-rebellion (the agent responds to the reaction of its decision from other agents).

In this section, we propose our framework in human induced interaction failure management for teleoperation tasks and explain the different stages involved with suitable examples. Lastly, in Figure 2, we point out the similarities and differences of our framework compared to the work done by Aha et al., and Peter Stone et al. mentioned above.

3.1 Framework

Using the example of ID behavior in guide dogs, we reverse engineered how the dog was able to decide and answer two vital questions as when to disobey and what to do after disobeying the handler's instructions, and divided the entire process into three stages as follows.

(1) Identification of failure trigger: Identify whether the action initiated by the user will lead to failure, given the local and global objectives, by reasoning on the available knowledge.
(2) Communication of feedback: Explaining to the user about the different processes of the system
(3) Mitigation of failure: Mediating the failure with an alternative course of action

3.1.1 Identification of failure trigger. The first step in managing failure is to identify the trigger/instruction that will lead to failure, and unlike physical failures, human failures are difficult to identify. The general approach towards identifying the failure is either to make the system learn the correct/ideal way of operation or to make the system learn the patterns of possible mistakes that a user might perform. Both of these methods are computationally expensive as each person might have different ways of working for the same task and cannot be generalized for other tasks. In our framework, the system predicts whether the given instruction by the human operator will lead to a failure event or not by reasoning if there is a conflict between the given current control sequence (local objective) and the predefined knowledge. Here, predefined knowledge consists of general norms of operation (e.g.: the speed of the robot must be low while turning or always keep a minimum distance between other objects in the scene, etc.) and global objectives (e.g.: reaching a location x or not harming the human, etc.) in the prioritized order. The idea is to check whether the local objective is to be disobeyed or not such that the global objective is satisfied depending on the general norms and environmental information.

3.1.2 Communication feedback. When the robot does not behave or operate the way that is intended, it leaves the operator with feelings of frustration or anger as the operator is unaware of the reason behind the actions of the robot. This action can range from slight delays to a complete halt of the process, depending on its processing time. Thus, updating the operator about the process the robot is currently doing is vital. This feedback improves the explainability of the system and reduces the negative perspectives the operator might acquire towards the system. In the teleoperation case, the only way of communicating back to the remote operator is by using notifications, in the form of visual and/or auditory signals, to grasp the operator's attention to the background process. In our framework, the feedback is the message that the given instruction by the operator is identified as a failure and is communicated back to the operator.

3.1.3 Mitigation of failures. The final stage is to effectively mitigate the failure. This stage makes use of other abilities that the robot can execute. The mitigation strategy could range from complete inaction to finding alternative ways to complete the task so that the global objective is not compromised. Knowledge about the abilities of the robot (e.g. steering in the direction where there are fewer obstacles present, etc.) and knowledge about the environment (e.g. location of objects in the scene of operation) is made available to the system, which helps in finding the alternative course of action. Feedback to the operator on the choice made to mitigate the failure is essential to prevent further conflicts and avoid unwanted assistance of the system to the user, which will result in degrading the perspective and trust towards the system by the operator.

3.2 Example use cases

To explain the functioning of the framework mentioned above in practical use cases, some hypothetical scenarios which correspond to different types of human induced interaction failure mentioned in Section 1 are detailed below.

Lapses: While teleoperating an unmanned aircraft, the user/operator is navigating the plane and suddenly gets distracted due to the "tunnel effect" (caused by driving in long stretches of path, which results in losing attention to the surrounding environment), and thus loses sight of the monitors and turns the joystick to a wrong direction that is opposite to the target location.

Deliberate violations: While tele-manipulating a pressure release valve in a nuclear power station, the user/ operator tries to open the valve above the threshold voluntarily, which could lead to a disastrous outcome.

Slips: During the task of delivering medicines remotely in hospitals, the operator accidentally (unintentionally) chooses to move forward while attempting to pick up the medicine package. Eventually, this results in damaging the medicine package located in front of the robot.

Mistakes: During an underwater teleoperation task for exploring and identifying artifacts, the operator comes close to an object. When the operator should slow down the speed, in order to avoid stirring the sea bed and fogging up the camera feedback, he/she chooses to intentionally approach the object at a higher speed (choosing a wrong method of operation).

The above-mentioned hypothetical failures are resulted due to the errors induced by humans, which could be prevented with the proposed framework. Here, failure is first identified by detecting the conflict between local and global objectives depending on the knowledge of the social norms of operation and the environment. The feedback will then be communicated to the operator and, finally, the decision on which tool to use to mitigate the failure. The above workflow is illustrated in Figure 3 as a generalized framework for any teleoperation scenario.

4 DISCUSSION AND FUTURE WORKS

"To err is human" and depending on the instance, the aftermath of the errors could be serious, especially in teleoperation applications. Factoring human errors and finding ways to manage them is a new and growing field of interest in HRI. Failure management systems have been existing in many forms such as assistance control (e.g.: lane keeping), and household (shared control vacuum cleaners) in automation. What differentiates the above control mechanisms from the proposed ID method is that the other control paradigms do not have external knowledge (global objectives, environmental conditions, or social norms of operation). They simply satisfy a particular condition without taking into account the overall picture. Systematizing this framework in terms of implementation as to

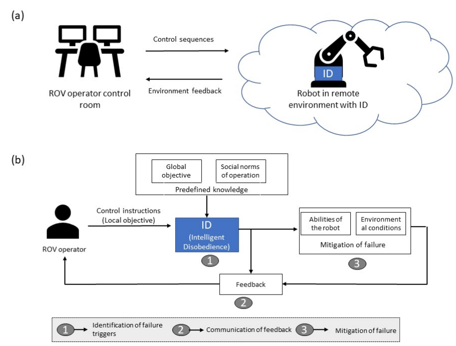

Figure 3: *(a)* Illustrates the general working of teleoperation where the operator is located in the control room and sends the control sequences through a communication medium to the robot, operating in the remote environment which is equipped with ID (Intelligent Disobedience) module *(b)* Elaborates the working of the proposed ID module where the three stages of ID are highlighted.

what methods can be utilized and how to practically implement this method with a proof-of-concept will be our future work. This poses challenges such as the global objectives and the general norms of operation having to be well defined. That will be an extensive task.

We aim to design this system to be human-centric i.e. the final control to override the system decision is given to the human, which helps the system to perform better in unstructured environments. However, it poses difficulty in managing deliberate human errors. Various questions, such as how fast the detection of failure should occur and how misclassification or non-classification of failures can be handled, are crucial to be analyzed further. We also acknowledge the ethical issues that surround this design, such as accountability, who takes responsibility for the decisions generated by this automated system, and to what extent we could rely on the automation systems to make decisions, which are difficult questions to answer with any given automation system. Posing these questions can help to design safe, reliable, and explainable systems for dealing with human induced interaction failures.

5 ACKNOWLEDGEMENT

This work has been supported by the Industrial Graduate School Collaborative AI and Robotics funded by the Swedish Knowledge Foundation DNR:20190128.

REFERENCES

[1] Scientific American, Nature America, and Scientific American. 2017. The CASE FOR Robot Disobedience Author (s): Gordon Briggs and Matthias Scheutz Source : Scientific American , JANUARY 2017 , Vol . 316 , No . 1 (JANUARY 2017), pp . 44-47 Published by : Scientific American , a division of Nature America , Inc . Stable. 316, 1 (2017), 44–47.

[2] Markus Bajones, Astrid Weiss, and Markus Vincze. 2016. Help, Anyone? A User Study For Modeling Robotic Behavior To Mitigate Malfunctions With The Help Of The User. *ArXiv* abs/1606.02547 (2016).

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

[4] 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.

[5] J. Carlson and R.R. Murphy. 2005. How UGVs physically fail in the field. *IEEE Transactions on Robotics* 21, 3 (2005), 423–437. https://doi.org/10.1109/TRO.2004.838027

[6] 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.

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

[8] Tomer Fine, Guy Zaidner, and Amir Shapiro. 2021. Grasping Assisting Algorithm in Tele-Operated Robotic Gripper. *Applied Sciences* 11, 6 (March 2021), 2640. https://doi.org/10.3390/app11062640

[9] Manuel Giuliani, Nicole Mirnig, Gerald Stollnberger, Susanne Stadler, Roland Buchner, and Manfred Tscheligi. 2015. Systematic analysis of video data from different human–robot interaction studies: a categorization of social signals during error situations. *Frontiers in Psychology* 6 (July 2015). https://doi.org/10.3389/fpsyg.2015.00931

[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] J-C. Laprie. 1995. Dependabe Computing and Fault Tolerance: Concepts and Terminology. In *Twenty-Fifth International Symposium on Fault-Tolerant Computing, 1995, ' Highlights from Twenty-Five Years'.* https://doi.org/10.1109/FTCSH.1995.532603

[12] Walterio W. Mayol-Cuevas. 2022. Rebellion and Disobedience as Useful Tools in Human-Robot Interaction Research – The Handheld Robotics Case. (2022). http://arxiv.org/abs/2205.03968 arXiv: 2205.03968.

[13] 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.

[14] Francisco-Angel Moreno, Javier Monroy, Jose-Raul Ruiz-Sarmiento, Cipriano Galindo, and Javier Gonzalez-Jimenez. 2020. Automatic Waypoint Generation to Improve Robot Navigation Through Narrow Spaces. *Sensors* 20, 1 (Jan. 2020), 240. https://doi.org/10.3390/s20010240 Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[15] 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.

[16] Samantha Reig, Elizabeth J. Carter, Terrence Fong, Jodi Forlizzi, and Aaron Steinfeld. 2021. Flailing, Hailing, Prevailing: Perceptions of Multi-Robot Failure Recovery Strategies. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Boulder CO USA, 158–167. https://doi.org/10.1145/3434073.3444659

[17] Robert Ross, Rem Collier, and G M P O'Hare. 2004. Demonstrating Social Error Recovery with AgentFactory. *3rd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2004)*, 3. https://doi.org/10.1109/AAMAS.2004.103

[18] 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

[19] Tachi Susumu and Komoriya Kiyoshi. 1984. Guide dog robot. In *Autonomous mobile robots: Control, planning, and architecture,*. 360–367.

[20] Hiroyuki Yasuda and Mitsuharu Matsumoto. 2013. Psychological impact on human when a robot makes mistakes. In *Proceedings of the 2013 IEEE/SICE International Symposium on System Integration*. 335–339. https://doi.org/10.1109/SII.2013.6776612