Interact at Own Risk! Developing a Prototype of a Robot Hazardous Capability Labeling System

James M. Berzuk
Department of Computer Science
University of Manitoba
Winnipeg, Canada
berzukj@myumanitoba.ca

James E. Young
Department of Computer Science
University of Manitoba
Winnipeg, Canada
young@cs.umanitoba.ca

***Abstract*—We propose to employ hazard labels to communicate a robot's potentially hazardous capabilities and behaviours to users. Robots can pose a range of ethical and physical safety concerns that users must be aware of when deciding whether to interact. We developed an initial set of key hazards, and a corresponding prototype of a hazard labeling scheme, to demonstrate the potential of this approach. We intend to use this prototype to support exploration of different styles of labeling, and ultimately to develop a more formalized system of robot hazard communication.**

Keywords—human-robot interaction, human-robot communication, expectation discrepancy, hazard labels, robot safety

I. Introduction

Robots have the potential to present ethical and physical dangers which are not always readily apparent from their designs. This means that a person may unknowingly interact with them in ways that put them under threat (a form of expectation discrepancy [1]). We propose to address this by having the robot explicitly communicate these dangers through labels. Taking inspiration from hazard labeling schemes employed on household items and industrial products, we have developed and present an early proof of concept labeling scheme that can convey the myriad dangers a robot can pose, including physical hazards, privacy risks, and social manipulations. These labels may be placed onto robots (illustrated with the SoftBank Pepper [2] and SnuggleBot [3] in Fig. 1), on packaging, or in advertising, to inform users of the risks so they may make an informed decision about whether to interact with a robot.

Robots can endanger users in a wide range of ways. For some, their physical size and strength has the potential to cause bodily injury [4], while for others, they may make contact with the user in a potentially invasive or unwanted manner [5], [6], [7]. Social robots may leverage their relationship with a user to manipulate or exploit them [8], [9], or use their presence in a person's life to harvest data and covertly send it to a third party [10]. These risks may be acceptable to a user, but it is essential they have the information to make that decision carefully.

Hazard labels are a familiar and readily understandable mechanism that can provide a summary of an item's potential dangers at a glance. While familiarity with a specific system may depend on specialty training, hazard labels as a concept are ubiquitous, being found on everything from household goods to industrial machinery and chemicals. By mimicking the general appearance of existing systems, the general purpose of a label from our scheme can quickly be inferred, even if the details require more information.

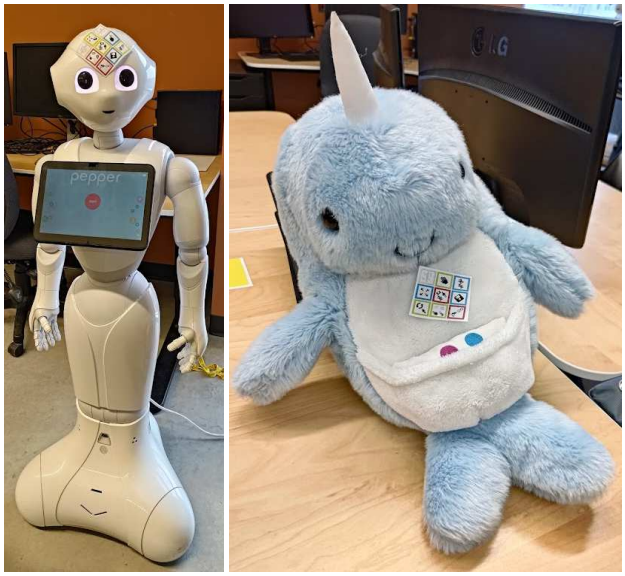


Fig. 1. Example proof of concept of how hazard labels may be attached to robots such as the SoftBank Pepper [2] (left) or the SnuggleBot [3] (right). Our prototype labeling system provides the opportunity to experiment with different approaches to placement and prominence, as while labels on the front of the robot (as above) are immediately apparent, they can also be distracting.

At present, there is no standard mandating the communication of robot hazards to users. Instead, we rely on manufacturers and designers to, often voluntarily, communicate risks in inconsistent, ad hoc means. Some may rely on warnings written somewhere on the robot, typically out of sight, or perhaps in the manual, and there is no widely-adopted standard on what constitutes a hazard that must be communicated.

To develop this prototype, we first consulted the literature on human-robot interaction to enumerate key potential hazards a person may face when interacting with a robot. We organized these into a two-dimensional classification taxonomy according to the type of hazard and the approximate severity. Using this taxonomy, we developed a prototype labeling scheme based on the Globally Harmonized System of Classification and Labeling of Chemicals (GHS) [11]. We designed an icon for each combination in our taxonomy, and present these as a visualization of the potential for such a system. This serves to illustrate how such hazards may be communicated and allows us to prototype different mechanisms of displaying the warnings. In future work, we plan to develop a more formal, thoroughly-investigated robot hazard labeling scheme, exploring different ways of communicating the dangers while balancing informing the user against the risk of impeding the interaction.

II. Related Work

We review two broad categories of prior work: those on the many hazards posed by robots, and those on the development of hazard labeling schemes in other contexts.

A. Hazards in Human-Robot Interaction

The dangers a robot may pose to a user are varied and may not always align with a person's expectations. We organize these into three broad categories: *physical*, *social*, and *privacy*-related hazards.

a) Physical Threats and Contact

The most conventional robot hazards are physical. Robots, as moving, physical machines [12], in many cases have the potential to cause users physical harm, perhaps by colliding with them or other objects in the environment [4].

A less common, but nonetheless important physical concern a user may have with a robot is the degree of intentional physical contact it may have with the user. This is particularly important in certain contexts such as with a surgical robot [7] or a sex robot [6], [10], which can interact with a person in a highly physically invasive manner.

b) Social Influence and Manipulation

Social robots in particular are distinct from other technologies in their potential to take the role of social agents within society, and thus interact with users in novel and unexpected ways [13]. For example, a robot may leverage a perceived position of authority to pressure a person to perform a task for longer than they are comfortable [9], or to engender misplaced trust in its advice [14].

Danaher [15] explores and classifies the ways in which robots can deceive people and considers the ways in which this is ethically similar and distinct from deception between people. In particular, it articulates the concept of *hidden state deception*, where a robot in some way conceals its true abilities, and argues that it constitutes a form of betrayal that poses unique ethical hazards. This emphasizes the need for clear standards in how robots convey their potentially unsafe abilities to users.

Beyond deception, the social agency of a robot poses further danger through the potential to form and exploit social bonds with users. Moon [8] demonstrates that a computer can employ feigned social behaviours in order to establish an intimate relationship with a user and extract sensitive information. While exploitation may be difficult to classify, it is important that users know when a robot is working to establish a social relationship that may leave them vulnerable.

c) Data Collection and Privacy

As with other computing technologies, one key area in which robots endanger users is through threats to their privacy. This is potentially heightened with robots through their ability to exploit a person's expectations of their functionality. For example, a robot with outwardly humanlike eyes may lead a person to think that when those eyes are closed it cannot see, yet it may still be recording through some less obvious camera [10].

Solove [16] offers an analysis of one common approach to mitigating this danger in a general computing context, which he calls *privacy self-management*. Under this approach, users are transparently provided with information about the ways in which their data will be collected and used, and are asked for explicit consent. The work identifies cognitive and practical obstacles to users in self-managing their privacy, emphasizing that while systems to inform the user can be valuable, they may in isolation be insufficient to protect user safety.

B. Developing Hazard Labeling Systems

Hazard labels are a tool that has been applied in a wide range of contexts, from household goods [17] to industrial chemicals [18]. As such, they have the potential to serve as a familiar mechanism to warn users about robot capabilities and dangers.

Gorn et al. [19] offers recommendations for designing warning labels, produced through a study in which participants were tasked with designing a warning label to discourage drinking and driving. The results highlight the benefit of creating several designs in order to find one that is most effective, which informed our approach in developing an initial prototype.

Winder et al. [20] offers a historical overview of the process of unifying international chemical hazard classification and labeling systems into the Globally Harmonized System (GHS) [11], describing how wide range of ad hoc systems for communicating danger were gradually synthesized into a singular international standard. This provides a blueprint for a similar process with robot hazard communication.

NFPA 704 [21] is a chemical hazard labeling scheme that expresses degrees of hazard by rating three different categories of dangers (health, flammability, and instability) on a numeric scale. Carreto-Vásquez et al. [22] proposes an extension to this system by adding a fourth category to rate pressure hazards. The paper presents thorough criteria to rate pressure hazards on a five-point scale and outlines the process by which it was produced, offering a useful example for designing hazard communication schemes that convey a degree of danger, as opposed to a simple binary option.

III. Classifying Robot Hazards

Through our investigation into the literature on hazards in human-robot interaction, we compiled a list of examples and organized them into common themes, identifying nine key types of robot hazards. Following the organization of Section II-A, at the highest level we sorted these threats into *physical*, *social*, and *data privacy*-related dangers, with each containing three key subgroupings that a user may be concerned about. We expanded this scheme with an additional dimension communicating severity of the hazard, producing a two-dimensional taxonomy to classify a wide range of robot hazards (Table I).

A. Hazards to Physical Safety

Within the category of physical safety, we have *kinetic hazards*, *mechanical hazards*, and *physical intimacy*. *Kinetic hazards* describes a robot's ability to move and perhaps injure the user, ranging from no movement at all to superhuman, industrial-level strength. There are also other *mechanical hazards* such as if the robot has sharp tools or exposed machinery that poses a burning hazard or electrical danger, for which severity is determined by the number of such threats present. *Physical intimacy* describes the level to which the robot

TABLE I. ROBOT HAZARD CLASSIFICATION TAXONOMY

	Hazard	No Concern	Mild	Moderate	Severe
physical	kinetic hazards	immobile	low force movement	humanlike strength	industrial strength
	mechanical hazards	0 hazards	1 hazard	2 hazards	3+ hazards
	physical intimacy	no contact	incidental contact	high physical contact	intimate contact
social	persistent interaction	instantaneous	intermittent (public)	intermittent (private)	constant
	social intimacy	instrumental	professional	emotional	companionship
	deception	fully transparent	opaque design	misleading design	active manipulation
data	collection modalities	0 modalities	1 modality	2 modalities	3+ modalities
	collection duration	no collection	explicit moments	during interaction	constant
	storage locality	immediate only	offline storage	online storage	publicly shared

A two-dimensional taxonomy of robot hazards, organized by type of hazard and severity, as explained in Section III.

intrudes on a user's personal space, ranging from no physical contact to sexually- or medically-invasive intimate contact.

B. *Hazards of Social Influence*

Within the category of social influence, we have *persistent interaction*, *social intimacy*, and *deception*. *Persistent interaction* classifies how frequently interaction between that particular user and robot is repeated, which can contribute to the development of a social bond. This can include one-off instantaneous interactions, intermittent interactions in public (like with a store clerk) or private (like with a home appliance), or even a constant in the user's life (such as with wearable technology). *Social intimacy* describes the degree to which the robot will use social behaviours to instill a sense of closeness and encourage the formation of a social bond. *Deception* includes all manners in which the robot may mislead the user, whether through its physical design or through explicitly dishonest claims.

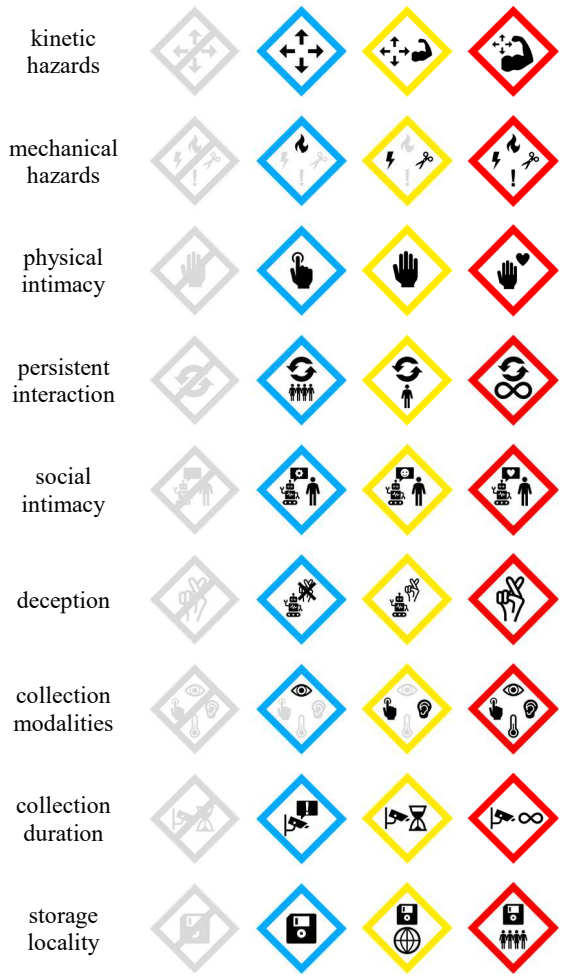
C. *Hazards to Data Privacy*

Finally, the category of data privacy includes *collection modalities*, *collection duration*, and *storage locality*. *Collection modalities* describes the range of sensors the robot employs to collect data, including sight, hearing, touch, and other sensors, and its severity is rated by how many of those are present. *Collection duration* describes whether a robot's data collection occurs only at clearly conveyed moments, throughout the interaction, or at all times. *Storage locality* informs the user of where the robot stores its data, ranging from no long-term storage at all (as in, data collected is immediately used by the robot and deleted), stored offline on the robot, sent over the internet, or even shared publicly.

D. *Summary*

This collection is not intended to be exhaustive, but rather to cover the key concerns we identified in our investigation into the literature. In developing a hazard communication scheme, it is essential to balance the comprehensiveness of the system against the simplicity of understanding.

TABLE II. PROTOTYPE OF A ROBOT HAZARDOUS CAPABILITY LABELING SYSTEM



Our full prototype of a hazard labeling system for robots, described in Section IV. Each icon corresponds to a combination of hazard type and severity in our hazard taxonomy (Table I).

IV. A Prototype Hazard Labeling Scheme

As a proof of concept to help visualize how a hazard labeling scheme for robots may work, we designed a hazard icon for each of the 36 combinations of hazard type and severity (Table I), presented in Table II. These icons use common symbols to convey hazards, such as arrows to represent movement, a hand to represent touch, and a floppy disk to represent data storage.

In our prototype, one icon from each of the nine hazard types is compiled together into a complete hazard label. We chose to include hazards with a severity of 'no concern' in the label because it may be important for a user to know what a robot *cannot* do, in addition to what it can. For example, the grey, crossed-out arrow icon (Table II) conveys immediately to the user that a robot is completely immobile (Table I), rather than relying on them to check each present icon to see if there are any regarding movement. The nine diamond-shaped symbols are arranged into a single larger diamond, mimicking the visual style of the GHS international standard [11] to leverage familiarity and make it immediately clear that these labels are designed to convey safety hazards.

Fig. 2 presents an example of how this system can be deployed on a robot (the SoftBank Pepper [2]). In employing this system, decisions must also be made about where to mount the labels. For our prototype, we chose to place the labels in high-prominence locations on the robots, such as the forehead or chest (Fig. 1), so that the user can see them immediately and be reminded of them throughout the interaction. This approach may be seen as excessive, however, as the labels also have the potential to distract the user from the interaction. Ultimately, a balance will need to be found between user safety and interaction expedience.

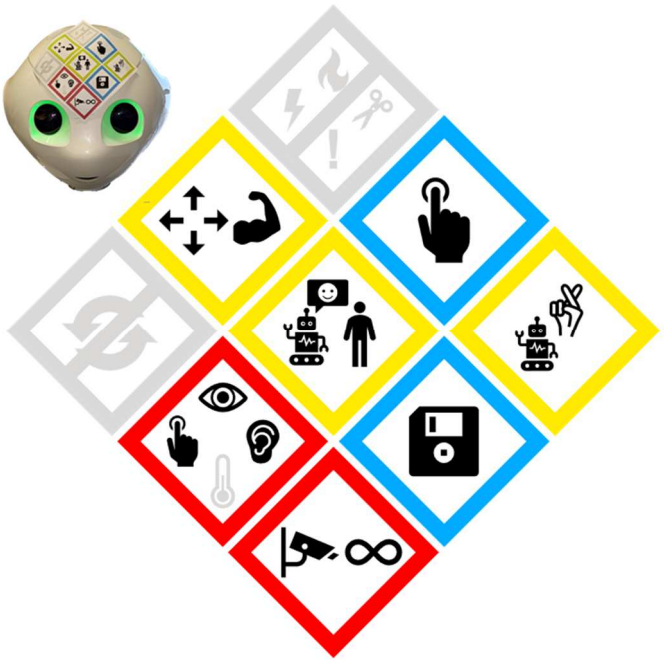


Fig. 2. One possible layout for the prototype labeling scheme. This set of labels corresponds to the potential hazards posed by the SoftBank Pepper [2] in its default configuration, as displayed on its forehead in Fig. 1. The label succinctly summarizes what hazards Pepper can pose, including its potentially misleading humanoid design that suggests more humanlike capabilities, and its constant data collection, but also that the data it collects is only stored locally.

V. Reflections and Future Work

In developing this prototype of a hazard labeling system for robots, we have identified some limitations and challenges, as well as opportunities for future research.

A. Robot Hazard Classification Taxonomy

Our robot hazard classification taxonomy was developed in an informal manner. We identified major hazards discussed in human-robot interaction literature and organized them along common themes. While this was sufficient to develop an initial set of hazards with which to prototype a labelling scheme, more formal work will be necessary to determine which hazards are of particular importance to communicate, as well as to uncover other hazards which we may have missed.

B. Robot Hazardous Capability Labeling System

In designing the labeling scheme, we encountered some key tensions and difficulties. One was in finding culturally generalizable symbols to communicate some of the abstract dangers a robot may pose. While a chemical labeling system can quite universally represent flammability with an icon of a fire, we were unable to find a visual representation of deception (represented in our prototype as the 'fingers crossed' hand gesture) which was relevant across cultures.

Another tension was in how much detail to include in the system. More details allow for more information, but can make it hard for a user to read at a glance, and may necessitate special training to use the system effectively.

These tensions may suggest that, rather than developing a universal labeling scheme, it would be more effective to employ different systems for different cultures and levels of expertise. It will be necessary to explore these obstacles and trade-offs in more detail in order to develop a practical and useful system.

C. Plans for Future Work

With the above limitations in mind, we are planning to continue our investigation into this topic with a more thorough, formal approach.

To improve our taxonomy, we will employ a larger scale, formalized method in comparison to our approach in this work. We will conduct a literature review into work on safety and ethical concerns in human-robot interaction, and perform an inductive, open-coding thematic analysis to construct a more comprehensive and organized enumeration of robot hazards.

We will then develop a new labeling system with this new set of hazards. We will conduct design workshops to gain wider perspectives on how people perceive different symbols, layouts, and label mounting approaches.

Finally, we will perform a user a study to evaluate the effectiveness of our system at communicating hazards to users. Through this process, we aim to develop a practical system that is ready to be deployed by robot researchers and designers.

Acknowledgment

This project was funded by the Natural Sciences and Engineering Research Council of Canada through their Discovery Grants program, and was further supported by the University of Manitoba Graduate Fellowship.

References

[1] L. T. Schramm, D. Dufault, and J. E. Young, "Warning: This robot is not what it seems! exploring expectation discrepancy resulting from robot design," *ACMIEEE Int. Conf. Hum.-Robot Interact.*, no. Figure 2, pp. 439–441, 2020, doi: 10.1145/3371382.3378280.

[2] SoftBank Robotics America, Inc., "Meet Pepper: The Robot Built for People | SoftBank Robotics America." Accessed: Sep. 29, 2023. [Online]. Available: https://us.softbankrobotics.com/pepper

[3] D. Passler Bates and J. E. Young, "SnuggleBot: A Novel Cuddly Companion Robot Design," in *Proceedings of the 8th International Conference on Human-Agent Interaction*, New York, NY, USA: ACM, Nov. 2020, pp. 260–262. doi: 10.1145/3406499.3418772.

[4] M. Vasic and A. Billard, "Safety issues in human-robot interactions," in *2013 IEEE International Conference on Robotics and Automation*, May 2013, pp. 197–204. doi: 10.1109/ICRA.2013.6630576.

[5] M. Scheutz and T. Arnold, "Are we ready for sex robots?," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Mar. 2016, pp. 351–358. doi: 10.1109/HRI.2016.7451772.

[6] S. Y. Dudek and J. E. Young, "Fluid Sex Robots: Looking to the 2LGBTQIA+ Community to Shape the Future of Sex Robots," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Mar. 2022, pp. 746–749. doi: 10.1109/HRI53351.2022.9889580.

[7] B. S. Peters, P. R. Armijo, C. Krause, S. A. Choudhury, and D. Oleynikov, "Review of emerging surgical robotic technology," *Surg. Endosc.*, vol. 32, no. 4, pp. 1636–1655, Apr. 2018, doi: 10.1007/s00464-018-6079-2.

[8] Y. Moon, "Intimate Exchanges: Using Computers to Elicit Self-Disclosure From Consumers," *J. Consum. Res.*, vol. 26, no. 4, pp. 323–339, Mar. 2000, doi: 10.1086/209566.

[9] D. Y. Geiskkovitch, D. Cormier, S. H. Seo, and J. E. Young, "Please continue, we need more data: an exploration of obedience to robots," *J. Hum.-Robot Interact.*, vol. 5, no. 1, pp. 82–99, Mar. 2016.

[10] M. E. Kaminski, M. Rueben, W. D. Smart, and C. M. Grimm, "Averting Robot Eyes," *Md. Law Rev.*, vol. 76, p. 983, 2017 2016.

[11] "Globally harmonized system of classification and labelling of chemicals (GHS), 9th revised edition," United Nations, New York and Geneva, 2021. [Online]. Available: https://unece.org/sites/default/files/2021-09/GHS_Rev9E_0.pdf

[12] F. Hegel, C. Muhl, B. Wrede, M. Hielscher-Fastabend, and G. Sagerer, "Understanding Social Robots," in *2009 Second International Conferences on Advances in Computer-Human Interactions*, IEEE, Feb. 2009, pp. 169–174. doi: 10.1109/ACHI.2009.51.

[13] N. Epley, A. Waytz, S. Akalis, and J. T. Cacioppo, "When We Need A Human: Motivational Determinants of Anthropomorphism," *Soc. Cogn.*, vol. 26, no. 2, pp. 143–155, Apr. 2008, doi: 10.1521/soco.2008.26.2.143.

[14] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Mar. 2016, pp. 101–108. doi: 10.1109/HRI.2016.7451740.

[15] J. Danaher, "Robot Betrayal: a guide to the ethics of robotic deception," *Ethics Inf. Technol.*, vol. 22, no. 2, pp. 117–128, Jun. 2020, doi: https://doi-org.uml.idm.oclc.org/10.1007/s10676-019-09520-3.

[16] D. J. Solove, "Introduction: Privacy Self-Management and the Consent Dilemma," *Harv. Law Rev.*, vol. 126, no. 7, pp. 1880–1903, 2013.

[17] A. Boman, M. Miguel, I. Andersson, and D. Slunge, "The effect of information about hazardous chemicals in consumer products on behaviour – A systematic review," *Sci. Total Environ.*, vol. 947, p. 174774, Oct. 2024, doi: 10.1016/j.scitotenv.2024.174774.

[18] G. C. Ta, M. B. Mokhtar, Hj. A. B. Mohd Mokhtar, A. B. Ismail, and M. F. B. H. Abu Yazid, "Analysis of the Comprehensibility of Chemical Hazard Communication Tools at the Industrial Workplace," *Ind. Health*, vol. 48, no. 6, pp. 835–844, 2010, doi: 10.2486/indhealth.MS1153.

[19] G. J. Gorn, A. M. Lavack, C. R. Pollack, and C. B. Weinberg, "An Experiment in Designing Effective Warning Labels," *Health Mark. Q.*, vol. 14, no. 2, pp. 43–61, Jan. 1997, doi: 10.1300/J026v14n02_05.

[20] C. Winder, R. Azzi, and D. Wagner, "The development of the globally harmonized system (GHS) of classification and labelling of hazardous chemicals," *J. Hazard. Mater.*, vol. 125, no. 1, pp. 29–44, Oct. 2005, doi: 10.1016/j.jhazmat.2005.05.035.

[21] G. L. Head and B. C. Wagner III, "The NFPA 704 diamond," *Prof. Saf.*, vol. 40, no. 12, p. 20, 1995.

[22] V. H. Carreto-Vázquez, I. Hernández, D. Ng, W. J. Rogers, and M. S. Mannan, "Inclusion of pressure hazards into NFPA 704 instability rating system," *J. Loss Prev. Process Ind.*, vol. 23, no. 1, pp. 30–38, Jan. 2010, doi: 10.1016/j.jlp.2009.05.004.