Identifying Ethical Hazards in Safety-Critical Systems: The Role of Creativity

Catherine Menon[1], **Austen Rainer**[2]

[1]Department of Computer Science, University of Hertfordshire

[2]School of Electronics, Electrical Engineering & Computer Science, Queen's University Belfast

Abstract Safety-critical systems can present a varied range of ethical hazards to users, operators and other stakeholders. Some of these hazards, such as a lack of fairness or transparency, have been discussed extensively in existing literature and appear in guidance documents and international standards. Others, such as cultural flattening, anthropomorphism, automation bias and systemic racial discrimination, are typically harder to identify and consequently harder to mitigate against. This paper presents an argument that creativity and collaboration play an essential role in ethical hazard analysis, and introduces a category of HAZOP-based techniques which can be used for a structured and creative discussion of such ethical hazards.

1 Introduction

A system's ethical hazards, like its safety hazards, can affect a diverse group of stakeholders, including users, owners, operators, and the general public. However, unlike safety hazards, there are no established and time-proven methods for identifying a system's ethical hazards, and no legal requirement to demonstrate that these are adequately mitigated. Moreover, although some standards exist which present general categories of ethical hazards (e.g BS 8611 (BSI 2023)) there are relatively few comprehensive evaluations of the potential ethical impacts of such systems. Because of this, ethical hazard identification is typically done, if at all, in an ad hoc manner which relies primarily on the effectiveness of the hazard identification team.

Moreover, ethical hazards present in many different ways and affect individuals, communities and demographics within society differently. When combined with the relative novelty of the field of ethical analysis of safety-critical systems, this means that not only is a diverse group of participants essential for ethical hazard analysis, but that creativity, collaboration and innovative thinking become essential characteristics.

In this paper, we present an argument for the importance of creativity in identifying ethical hazards, focusing on systems which incorporate AI. We also introduce X-HAZOP, a category of HAZOP-based techniques which help to facilitate and foster such creativity, and present the results of multiple workshops conducted to firstly assess the effect that X-HAZOP methodologies have on creativity, and to secondly determine whether these methodologies can be effectively used within the context of ethical hazard analysis of safety critical systems.

In Section 2 we describe the background and related literature. Section 3 introduces X-HAZOP as a HAZOP-based category, while Section 4 describes the workshops conducted using these techniques and summarises the results. Section 5 presents a discussion and analysis, and Section 6 concludes.

2 Background

Existing literature such as (BSI 2023; IEEE 2018; Leslie, 2019) considers ethical hazards in AI-informed systems from the perspective of systems design, with a number of characteristics such as transparency, fairness and freedom from bias being identified as ethically desirable. However, the majority of these works do not specifically consider the identification and mitigation of ethical hazards, nor do they present any discussion of a process or methodology which could be used. One exception here is (BS 8611) which presents a list of ethical hazards which may be specifically relevant to assistive robots, including the ethical hazards of inappropriate trust in the robot, deception, anthropomorphisation and robot addiction. However, there is little information about how these hazards may be generalised to other safety-critical domains. One of the few methodologies for ethical hazard identification is that given in (Winfield et al, 2022), which considers a hypothetical smart robot toy resembling a Furby; nevertheless this field remains largely under-explored.

Similarly, although HAZOP itself is a well-established technique within safety analysis (BSI 2016), there are relatively few studies exploring how this might be extended to other analysis domains, including ethical analysis. The concept of

[1] c.menon@herts.ac.uk

[2] a.rainer@qub.ac.uk

targeted, or "trigger" questions to encourage creative facility is in itself well-researched, with (Bloom, 1956) and subsequent research (Reynolds, 1996) (Goth, 2011) identifying how this form of structured assessment enhances both collaboration and creativity. However, such studies are based around the use of creativity for fiction, which is of necessity less directed or structured than the process of identifying ethical hazards within a safety-critical system.

With this in mind we propose the X-HAZOP family of techniques for ethical hazard analysis. These bridge the gap between the "pure" creativity of fictional critique methods, and the high-level overview of desirable ethical system characteristics identified within existing standards. In this way, we propose that X-HAZOP is a reproducible technique to identify ethical hazards while at the same time facilitating the creativity and collaboration which is essential to adequately assure that all such ethical hazards have been addressed.

3 X-HAZOP methodologies

The X-HAZOP family of methodologies is informed by, and draws upon, the HAZOP methodology for hazard identification in safety-critical systems. The X-HAZOP family consists of a set of structured, systematic processes by which guidewords can be used to assess a system, where "system" in this context includes both technological and non-technological entities: i.e. systems, peoples and processes. The X-HAZOP family extends HAZOP by placing the focus of the analysis specifically on the creative discussions which emerge from the guidewords, rather than the rote use of the guidewords themselves.

In this section we present two of the X-HAZOP methodologies: CHAZOP (Creative HAZOP), designed to assess the narrative behind a system, and EHAZOP (Ethical HAZOP), designed to assess the ethical properties of a system.

3.1 CHAZOP

CHAZOP (Creative HAZOP) was defined to assess the effectiveness of HAZOP-based methodologies in fostering and encouraging creativity. CHAZOP is a narrative-based approach which may be used in constructing the kinds of narratives seen in user stories, vignettes[3], persona definition[4]s and also in the process of narrative approaches to software design. In defining CHAZOP as below, we focused on its use in fiction, as this is a natural environment to assess the method's contribution to creativity.

CHAZOP makes use of a facilitated creative discussion founded on the application of pre-defined *guide-words* to components of the narrative (referred to as *narrative components*). Within our instantiation of CHAZOP the *narrative components* are defined to be:

- Plot - the structure and relationship of events within the narrative
- Setting - the environment (physical, thematic etc.) wherein the narrative takes place
- Character - the persons or other entities undertaking *plot* action or taking part in the narrative
- Voice - the overall effect on the reader's perception of literary devices including syntax, vocabulary, metaphor and imagery
- Theme - the development of underlying ideas, perspectives or resonances not necessarily directly stated in the narrative
- Linguistic creativity: the language, word and style choices and sentence structure
- Friction: the tension between *characters*, or the obstacles placed in a character's way
- Structure: the perspective, tense, composition and organisation of the narrative
- Point of view: the perspective from which the narrative is told

There are likely to be additional narrative components for any given narrative and teams should add any that they consider useful. Depending on the purpose of the narrative, additional narrative components might include imagery, suspense, humour or allusion.

As with HAZOP, CHAZOP makes use of *guide words*. CHAZOP guide words are presented in Table 2 and are intended to draw out collaborative opinions on the opportunities for improvement within a narrative.

Not all the guide words of CHAZOP will be relevant to every narrative. Even where relevant, any one guide word may not indicate an undesirable characteristic: for example it may be that the subsequent discussion following the application of REVERSE to the narrative component PLOT reveals that an unreliable narrator has deliberately been chosen by the author.

[3] A vignette is a short narrative account or description of an experience with the system

[4] A persona is a description of somebody who may feasibly interact with a system

Table 1. CHAZOP guide words

Guide word	Meaning
NOT ENOUGH	Insufficient emphasis, clarity or focus on this aspect of the narrative
TOO MUCH	Over-emphasis on this aspect of the narrative, at a cost to other aspects or reader attention
EARLY	Information conveyed to reader earlier than the writer's assumed intent
LATE	Information conveyed to reader earlier than the writer's assumed intent
NEVER	Information never conveyed to reader with sufficient clarity
AS WELL AS	Side effect or information conveyed to reader in addition to the writer's assumed intent
REVERSE	Opposite of the writer's assumed intent
MORE	Effect of increasing the specified narrative component
LESS	Effect of decreasing the specified narrative component
BEFORE	Effect of some aspects of a narrative component being encountered by the reader before others
AFTER	Effect of some aspects of a narrative component being encountered by the reader after others
SAME	Information about this narrative component remains the same through-out the piece
DIFFERENT	Information about this narrative component is different or inconsistent with other information gathered by readers

When conducting CHAZOP the team must work systematically through the guide words and narrative components. Creativity is essential to maximise the benefit of CHAZOP. For example, any of the following illustrative discussions might arise when applying the guidewords:

- NOT ENOUGH – PLOT: "I don't think enough happens"
- EARLY – PLOT: "What if we found out about the character's motivations earlier?"
- MORE – THEME: "I found the theme to be under-developed and would like more of this"
- REVERSE – THEME: "What if the opposite (literary, political, moral…) perspective were presented?"
- AFTER – CHARACTER: "What if we found out about the character's motivations after an alternate event?"
- DIFFERENT – VOICE: "I would have connected more if the voice had been simpler or clearer"
- AS WELL AS – VOICE: "What if this narrative were told from multiple narrative perspectives? Would that provide additional useful information?"

CHAZOP may be used in one of two modes: either to help a team identify and describe opportunities for improvement in a narrative or to help the same group analyse a narrative which is unfinished, in order to find how it might usefully be progressed.

3.2 EHAZOP

EHAZOP (Ethical HAZOP) was defined to assist a team in asking structured, guided "what if" questions to determine whether any ethical hazards are manifested by the delta between stakeholder expectations and system implementation. Ethical hazards are always considered from the perspective of the stakeholder themselves – for example, the perception of an ethical hazard such as bias can in itself represent ethical harm, whether or not this hazard eventuates (BSI, 2023) – and so as such we consider the guidewords relative to a stakeholder's expectations of the system. In the instantiation below, we focused on EHAZOP's application to an assistive robot, as we had the infrastructure (see Section 4) to support this during the workshops.

EHAZOP makes use of a facilitated discussion founded on the application of pre-defined *guide words* to physical components of the system, functions of the system, characteristics of the system or combinations thereof: these are referred to as *system modules*. Within our instantiation of EHAZOP (to an assistive robot) the *system modules* are defined to be:

- Specified robot functions
- Specified robot characteristics, these being:
    - Robot non-functional requirements
    - Aspects of the specified robot physical design
    - Extent of robot autonomy

There may be additional system modules for any given system, and teams should tailor these accordingly.

As with HAZOP, EHAZOP makes use of guide words. EHAZOP guide words are presented in Table 2 and are intended to draw out collaborative opinions on the potential ethical hazards.

Table 2. EHAZOP guide words

Guide word	Meaning
EARLY	This characteristic or function of the system occurs or is encountered earlier than the user expects
LATE	This characteristic or function of the system occurs or is encountered later than the user expects
NEVER	This characteristic or function of the system never occurs or is encountered despite being expected by the user
AS WELL AS	This characteristic or function of the system is performed or encountered in addition to a different one expected by the user
REVERSE	Opposite of the writer's assumed intent
MORE	This characteristic or function of the system is more or increased from that expected by the user
LESS	This characteristic or function of the system is less or decreased from that expected by the user
OPPOSITE	This characteristic or function of the system is the opposite of that expected by the user

EHAZOP guide words are used in combination with the system modules. For example, any of the following illustrative discussions might arise when applying the guidewords:

- What if this function were provided ⟨ EARLIER ⟩ than the user expects?
- What if this function were provided with ⟨ LESS ⟩ ⟨AUTONOMY ⟩ than the user expects?
- What if the robot had the ⟨ OPPOSITE ⟩ ⟨ physical design ⟩; how would this affect user expectations of each function?

It is important to note that, as is the case for HAZOP, not all the EHAZOP guidewords will be applicable to each function or characteristic.

4 Workshops and validation

We conducted two preliminary CHAZOP workshops, and two preliminary EHAZOP workshops. For both CHAZOP and EHAZOP the workshops were intended to serve only as a proof of concept, with full validation of the X-HAZOP methodologies to follow[5].

4.1 CHAZOP workshops

Both pilot study workshops were carried out at the Crescent Arts Centre in Belfast, using the same experimental process. Participants were recruited using social media advertising, personal outreach by Crescent Arts in accordance with GDPR regulations, and advertisement on the Crescent Arts homepage.

After obtaining consent, participants were provided with an overview of the CHAZOP process and a high-level introduction to its aims. All participants had been asked to submit a written narrative (piece of fiction) prior to the workshop. These were anonymised and distributed to all participants prior to the workshop.

Participants conducted a CHAZOP process on each of the written narratives in turn, facilitated by staff from Queen's University Belfast and the University of Hertfordshire. Each written narrative was approximately 5000 words long, and the CHAZOP process lasted approximately one hour for each.

[5] The CHAZOP workshops were both approved by the University of Queen's University Belfast Faculty of Engineering and Physical Sciences' Ethics Committee under protocol number EPS 22_297, and a full report is provided here. The EHAZOP workshops were approved by the University of Hertfordshire's Health, Science, Engineering and Technology Ethics Committee, and a summary of findings only is provided here.

4.1.1 Post-study questionnaires

Following both of the workshop, participants were asked to complete an anonymous post-study questionnaire, requesting both qualitative and quantitative feedback.
The quantitative feedback for both workshops included the following questions, in order to determine the effect that CHAZOP had on participants' perceived creativity:

- Participants were asked to give a numerical score of whether they considered CHAZOP useful for incomplete and for complete narratives (0 = not at all, to 5 = very useful)
- Participants were asked how helpful they found the CHAZOP process to them personally as a writer (0 = very unhelpful; 5 = very helpful)
- Participants were asked how helpful they found the CHAZOP process as a team participant assessing a narrative (0 = very unhelpful , 5 = very helpful)

4.1.2 Results

As this was a preliminary study, with correspondingly low participant numbers (11 participants across the two workshops), no statistical significance between conditions and questionnaire responses was expected. Nevertheless, there were indicative trends to support our hypothesis that CHAZOP is perceived to be of creative benefit when used for analysis of a narrative.

Participants across both workshops felt strongly that CHAZOP was of creative benefit (82% - 90% agreement). Participants also considered that CHAZOP would present greater benefit to inexperienced narrative writers than experienced writers. Amongst such inexperienced writers, there was 100% consensus that CHAZOP would be creatively helpful. Amongst professional narrative writers, there was a conclusion that CHAZOP nevertheless provided moderate creative benefit to themselves (mean value 0.5).

Qualitative responses were also sought from the participants via free-text feedback. These were overall very positive, with the following comments recorded on the questionnaire:

- "Very helpful to enhance creativity - a sense of structure can inspire new ways of thinking"
- "[I]t leads to inspiration and helps re-energise creative process"
- "I will be teaching creative writing undergrads this term and I will use it"
- "I think it offers a useful language for critiquing that (to some extent) removes the element of judgement"
- "This is an excellent roadmap for giving beginner writers the tools to evaluate their own and others' writing"

4.2 EHAZOP workshops

Two EHAZOP workshops were performed, the first being a proof of concept workshop carried out in the University of Hertfordshire's Robot House involving five participants with backgrounds spanning ethical analysis, safety, architecture and narrative analysis. The system used for ethical analysis in the workshop was Ari, an assistive robot equipped with movement control, a touch screen and gaze direction for human interaction (Pal Robotics, 2024).

Fig. 1. EHAZOP workshop

The second EHAZOP workshop was conducted with school students, again with the focus being an assistive robot such as Ari.

4.2.2 Results

The results of this second EHAZOP workshop have not yet been fully analysed and are therefore are not presented here, but for the first EHAZOP workshop, participants identified a number of ethical hazards associated with the assistive robot, including:

- Lack of privacy, lack of informed consent, dehumanisation, robot addiction, erosion of confidence (guideword: MORE)
- Loss of trust, lack of respect for cultural diversity and pluralism, lack of associative control (where the user's mental associations with performing an activity are altered due to the robot's reception of it) and deception (guideword: OPPOSITE and OPPOSITE combined with AUTONOMY)

Although time pressures did not permit the full spectrum of guide words and system modules to be iterated, nevertheless even in this abbreviated form, participants were able to identify novel ethical hazards that are not included in typical enumerations of these (BSI, 2023). These included *erosion of confidence* and *lack of associative control.*

Participants also considered that a number of these hazards would not have been identified without the structured and facilitated creativity fostered by use of EHAZOP. One of the most compelling takeaways from the EHAZOP workshop is the value of the ensuing discussion after the guidewords: all participants considered that this was the most useful part of the workshop. Participants also took the initiative to propose alternate functions, system modules and combinations thereof which would help mitigate the ethical hazards.

5 Discussion and Analysis

Both the CHAZOP and EHAZOP workshops demonstrated that these techniques enhance perceived creativity in structured analysis and, moreover, can be used effectively to conduct ethical hazard analyses of safety-critical systems. Although no statistically significant conclusions can be drawn from either methodology workshops, participants in all were uniformly positive about the use of the methodologies.

Moreover, we identified for both CHAZOP and EHAZOP that, although participants greatly appreciated the presence of a facilitator, they all considered that the most helpful results were obtained in the discussion *after* using the guidewords, i.e. in the more creative part of the workshop. The associative links between guide words, narrative form and ethical hazards is extremely valuable, and thus we would suggest that the facilitator role should transition from facilitator to amanuensis in the latter part of the workshops.

We also found that not only was creativity enhanced during the workshops, but collaboration was also fostered. Participants self-reported that they were able to consider suggestions for their narratives – or suggestions for amendments to their identified ethical hazards – and build on these as part of a team. In addition to this, the majority of the participants in both workshops also identified that the narrative improvements and ethical hazards which were suggested would be unlikely to have been identified without the use of the X-HAZOP methodology under analysis.

6 Conclusions

The X-HAZOP family of methodologies are a category of techniques which draw on HAZOP, and which can be used to facilitate and foster enhanced creativity and collaboration in ethical hazard analysis. In this paper we have discussed why creativity is such an essential part of ethical analysis, and presented some of the obstacles towards adoption of a strictly process-based, standards-dependent approach.

We have also introduced two of the X-HAZOP methodologies: CHAZOP (Creative HAZOP) and EHAZOP (Ethical HAZOP). Both of these methodologies use pre-defined guidewords; in the case of CHAZOP these are applied to a narrative such as a user story or vignette, while in the case of EHAZOP these are applied to system components or functions.

We have described workshops which took place to partially validate both CHAZOP and EHAZOP and to further refine these techniques. Participants in both of the workshops uniformly considered the techniques helpful to fostering creativity and in the case of EHAZOP were additionally able to identify multiple ethical hazards associated with an assistive robot which are not present in current standards containing enumerated lists of such hazards.

We propose to build on this work by defining further methodologies for the X-HAZOP family, focused specifically around the ethical analysis of safety-critical systems involving AI. Such methodologies would need to integrate the creativity demonstrated by CHAZOP, along with specific targeted discussion points to enable the potential ethical impacts of AI to be fully discussed and realised. In order to support this, we also propose to fully analyse the results from the EHAZOP workshops to determine how to ascertain the needed diversity of composition of EHAZOP teams, in order to ensure that sufficient diversity of perspectives is achieved in ethical analysis using these methodologies.

References

Bloom, B. (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals. In Handbook 1, Cognitive Domain. New York: David Mckay.

British Standards Institute (2016). BS 61882: Hazard and Operability Studies – Application. https://knowledge.bsigroup.com/products/hazard-and-operability-studies-hazop-studies-application-guide?version=standard

British Standards Institute (2023). BS 8611: Robots and robotic devices. Guide to the ethical design and application of robots and robotic systems. https://standardsdevelopment.bsigroup.com/projects/2022-00279#/section

Goth, J., Ha, E., Lester, J. (2011). Towards a Model of Question Generation for Promoting Creativity in Novice Writers. In AAAI Fall Symposium Series, pp 23 – 26.

IEEE (2018). Ethically Aligned Design. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the respons ble design and implementation of AI systems in the public sector. The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529

Pal Robotics (2024). Ari: The Social and Collaborative Robot. https://pal-robotics.com/
robots/ari.

Reynolds, T., Bonk, C. (1996). Computerized Prompting Partners and Keystroke Recording Devices: Two Macro Driven Writing Tools. In Educational Technology Research and Development, Vol 44(3), pp 83 – 97.

Winfield, A. et. Al. (2022). Ethical Risk Assessment for Social Robots: Case Studies in Smart Robot Toys. Towards Trustworthy Artificial Intelligent Systems. Intelligent Systems, Control and Automation: Science and Engineering, vol 102.