UNIVERSITY OF TWENTE.

Academic Research Skills
Conference Proceedings

internal edition

Safe integration of cobots enjoying artificial intelligence in healthcare environments

Bart Hendriksen, Raymond Harhuis, Mohammad Rajabali Nejad

Pre-master track: ME

Submission date: January 28, 2020

As artificial intelligence gains more importance in various aspects of society, the safe relationships between the system, its environment and humans become of more and more importance. In order to give guidance to the safe integration of these AI enjoying cobots, design constraints and safety indicators are defined in this paper, but first the question is answered whether this possible at all. Using a stakeholder analysis, the safety cube theory and a N-squared diagram, these design constraints and safety indicators are derived. The notion of safety is extended in this paper from human safety to safety for humans, the system and its environment. In conclusion, safe integration is possible, when the constraints and indicators are considered, and more elaborate laws are created regarding AI technology.

Artificial intelligence, Cobots, Healthcare, Safety Cube Theory, safe integration

1. Introduction

Artificial intelligence (AI) is a theme that emerges to be more and more important in the society of the twenty-first century. Machines that enjoy AI are becoming ever more efficient in their jobs. Working through vast sets of data is a challenge for a human, but a neural network can be taught to work its way through this data with relative ease. As it processes more data, it grows smarter. Whereas machines were invented to make the life of humans easier and be under the command of humans, the question now arises whether machines can control the lives of humans autonomously and to what extent that is acceptable. To answer the question whether this technology is inherently safe is simply impossible without a vast research. Based on the gap found in literature, the question one must ask is whether it is safe to use AI systems in healthcare cobots in the specific case of assistive cobots. As the focus lies thus on safe integration in this area, one must first define safety.

Safety is '*freedom from those conditions that can cause death, injury, occupational illness, or damage to or loss of equipment or property, or damage to the environment*' [1] The indicators for a safe performance of healthcare cobots then are derived from this statement: 'the cobot must not harm humans, environment and other materials.' So, the main indicator for a safe performance of assistive healthcare cobot is that the machine does no harm to anything it is in contact with or itself.

The current method for the integration of cobots and other machines is the ISO standard. This is produced as a guideline for machines and the integration of these machines to ensure a safe working. This standard does not include specific parts for the integration of artificial intelligence. Because there are no alternatives, industries use this standard at current time. Final analysis in this paper also aims to find whether this standard gives sufficient boundaries for a design.

This comes down to answering the question: *'How can it be ensured that cobots who enjoy AI are safely integrated in a healthcare environment.'*

2. Methodology

This paper aims to define a strategy by which cobots in healthcare technology (explicitly in the context of assistive cobots in care-homes or other supportive facilities) can be safely integrated. Note that a cobot in this context is defined to be a robot designed and intended for direct physical interaction with humans. To obtain this, the cobot enjoying AI is analysed using the safety cube theory using stakeholders as an input, the interaction between humans, the system of interest (SoI) and their environment are analysed. Based on this, a N-squared diagram is derived. Then, using a fishbone analysis, design constraints and safety indicators are defined.

3. Safe integration of healthcare cobots

To find an approach to safely integrate these cobots, safety indicators must be derived. These can be divided in leading (or active) and lagging (or passive) indicators. Leading indicators are (the active monitoring of) preceding measures in order to prevent safety hazards from occurring, while lagging indicators are reactive in nature. One could think of number of times an emergency stop is used to be a lagging indicator and actively training people to be a leading indicator. To derive this, a systematic approach as described in [§2. Methodology] is used.

3.1 Stakeholders and users of healthcare cobots

To analyse stakeholders, one must look at possible life cycles of a product, in this case an assistive healthcare cobot operated mainly by a self-learning AI. Healthcare cobots are not (yet) mass produced. Thus, a general life cycle will apply to a vast majority of the cobots. Opposed to Rogers' bell curve, the product is judged by its technical life cycle. This can be summarized in eight stages, not necessarily in this order: conception, development, production, sales, (installation), operation, maintenance, disposal. In [Table 1] the different stakeholders with their interest are shown in each of the phase of the cycle. Note that this list is rudimental, as interests are in practice very specific and more complex than in theory. Even though the list is not necessarily in chronological order, it is cumulative in causal sense, meaning that stakeholders are mainly interested in a specific phase and at least have some interest in the following phases. The designing, engineering, manufacturing, installation and maintenance companies can be the same company.

When not specifically mentioned, it is implied that all stakeholders should have safety as an interest. This list of stakeholders is not pretending to be complete and yet already

contains twenty members. This results in a very lengthy analysis in the N-squared diagram. Since this paper aims at safe integration, one can omit several stakeholders that do not actively influence the safety aspects in the operation of the cobot.
The five principal stakeholders can be found in the N-squared diagram.

Phase	Stakeholders	Interest and influence
Conception	Designing personnel	Design an attractive product; large influence on design
	Engineering personnel	Design a durable product; large influence on design
	Patients	Desire a safe appliance; medium (indirect) influence on design
Development	Governing bodies	Desire a safe product; very large influence through laws
	Engineering company	Desires a durable and easy to sell product; large influence on product.
	Software developers (AI)	Desire a cobot operated by AI; large influence on operation of product.
	Customers (statement of needs)	Desire a durable, trustworthy and safe product for the least possible amount of money; medium influence on design
Production	Production personnel	Desire easily producible product; low influence on design.
	Production methods	Limit the design possibilities; large influence on the design.
Sales	Dealers	Desire a sellable product; low influence on design.
Installation	Installation personnel	Desire a simple product; low influence on design.
	Medical staff	Desire an easy to use system that does part of their job; medium influence on design.
	Care home staff	Desire an easy to use system that does part of their job; medium influence on design.
	The AI itself	Limits the capability of the cobot, if designed properly, wants to assess risks; has low influence on design.
	Nurses	Desire an easy to use system that does part of their job; medium influence on design.
Operation	Visitors of patients	Desires a system that safely handles their family members/friends; low influence on design
	Volunteers in medical/care home facilities	Desire an easy to use system that does part of their job; low influence on design.
Maintenance	Maintenance personnel	Desire easy to maintain system; low influence on design.
Disposal	Recycling plant personnel	Desire easy to dismantle system; medium influence on design.
	Waste disposal facility personnel	Desire easy to dismantle system; medium influence on design.

Table 1.: Stakeholders, their interests and influences

3.2 The relation between the healthcare cobots, humans and environment.

The relation between the cobot, stakeholders, users and the environments the cobot operates in can be expressed using the safety cube theory [2]. This describes the hierarchical and behavioural aspects of the relations between these factors as is shown in [Table 2]. Based on this general relationship, the specific relationships between the individual members along the diagonal of the N-squared diagram can be derived.

	Human	*System*	*Environment*
Human	*Medical personnel, Patients, Hospitals*	*Input from medical personnel and patients*	*The hospital, care home, where the cobot is deployed.*
System	*Input from medical personnel and patients*	*Supporting medical personnel and patients with everyday tasks.*	*Inputs from the surrounding of the hospital or care home*
Supersystem or Environment	*Input from medical personnel and patients*	*Inputs from the surrounding of the hospital or care home*	*Regulations for the cobots, cooperating with other cobot.*

Table 2.: General relationships by the safety cube theory

3.3. Applying the N-squared method

The information above can be applied using the N-squared method. In this paper, the authors chose to keep the diagram limited to the twelve most important aspects of the integration of the SoI in their eyes. This results in 132 different relations or influences as shown in [Table 3.]

3.4 Failure modes & current integration methods

Fault modes (also known as failure modes) are states in which a system does not operated as expected or wanted, often with negative consequences for the stakeholders or the machine itself. Faults can be environmentally induced (including human errors) or inherent to a system. Analysis in engineering practice (mechanical fault modes) is usually executed using methods like the failure mode and effects analysis (FMEA) or the variant adding criticality that was used by NASA (FMECA)[3]. Use of these methods can be a secondary step to safe integration of AI cobots in healthcare technology. The focus of this paper will be more on the inherent fault modes to AI systems. The relations between and integration among SoI/human/environment were already established using the N-squared method. Fault modes can occur in, but are not limited to, any of the 132 relationships or influences in the diagram. A cable defect or a non-functioning network adapter can result in interrupted communication between most of the elements on the diagonal. A staff member's or even the AI's inability to activate the emergency stop is an example of a fault mode in the human-system integration. On human-environment integration, a communication error induced by the cobot could occur between a patient or staff member and the network. A fault mode could then be the inability of the patient to operate the network or the network to monitor the patient. The same holds for system-environment integration. The absence of properly applicable laws can cause major problems in later stadiums of operation and bring large economic costs with it.

	Human				System				Environment			
Human	**Patient**	Relies on	Needs support of	Provides input for the maintenance staff.	Sensors detect movement and need of patients.	-	Can override AI and communicate with it.	Patient may operate the emergency stop.	-	Physical environment is designed for patient care.	Patient may alter behaviour of cooperating systems.	Patient may provide input for network.
	Cares for	**Medical personnel**	Needs support of	Provides input for the maintenance staff.	Sensors detect movement and input from personnel.	-	Can override AI and communicate with it.	Personell may operate the emergency stop.	Input from personnel could improve laws.	Physical environment is designed as safe work environment	Personnel may alter behaviour of cooperating systems.	Personnel may provide input for network.
	Daily care tasks	Supports	**Care home staff**	Provides input for the maintenance staff.	Sensors detect movement and input from staff.	-	Can override AI and communicate with it.	Staff may operate the emergency stop.	Input from staff could improve laws.	Physical environment is designed as safe work environment	Staff may alter behaviour of cooperating systems.	Staff may provide input for network.
	Maintenance staff guarantee prolonged safe use of cobot.	Maintenance staff provide safely operable system.	Maintenance staff provide safely operable system.	**Maintenance staff**	Maintenance staff install, calibrate and repair sensors.	-	Can override AI and communicate with it.	Maintenance staff install, calibrate and repair emergency stop.	Input from maintenance staff could improve laws.	Physical environment is designed as safe work environment	Maintenance staff may alter behaviour of cooperating systems	Maintenance staff may alter network and provide input for the network.
System	-	-	-	Maintenance staff can extract data from sensors.	**Sensors**	-	AI can extract data from sensors.	Sensors can log the use of emergency stop.	-	-	-	-
	Motors provide power to handle patient.	-	-	Motors are more prone to damage than nonmoving parts.	Motor actions form input for sensors.	**Motors**	Motors constrain for example available force for the AI to use.	-	-	-	-	-
	AI can assist patient using the cobot	AI can do part of the job of personnel	AI can do part of the job of staff	AI can provide input for maintenance staff.	Sensors can measure AI actions.	AI needs motors to operate cobot.	**Artificial intelligence**	AI can shutdown system using emergency stop.	AI requires new laws.	AI interacts directly with physical environment.	AI can cause cooperating systems to alter behaviour.	AI communicates with network and can actively influence it.
	Guarantees safe patient handling.	Guarantees safe work environment.	Guarantees safe work environment.	Guarantees safe work environment.	AI or manual override could disable sensors (undesirable)	AI or manual override can disable motors.	Manual override emergency stop can shutdown AI control.	**Emergency stop**	-	-	One emergency stop could should down multiple systems.	Emergency stop occurence needs to be communicated
Environment	No direct laws apply, but patient is subject	Arbo is applicable	Arbo is applicable	Operate according to arbo	Subject to 2006/42/EG	Subject to 2006/42/EG .	Can restrict AI use and employabillity.	Emergency stop is required by law.	**Laws and regulations**	Constraints on physical objects.	Machinery directive and other applicable standards.	AVG and other restrictions on building networks.
	Is handled in physical environment.	Need to work in physical environment.	Need to work in physical environment.	Need to work in physical environment.	Physical environment passively provides input for the sensors.	-	Physical environment provides input for the AI.	Constrains the possible uses of emergency stop.	Different applicable laws and regulations.	**Physical environment**	Physical limits constrain possible application of systems.	Network is finetuned to physical space.
	Patient is handled by the cooperating systems.	Can alter behaviour of personnel based on output data.	Can alter behaviour of staff based on output data.	-	Cooperating can use sensors as a means of communication	-	Can actively communivate with control system	Cooperating systems may activate emergency stop	Advances in technology alter standards.	Systems occupy physical space.	**Cooperating systems**	Network is adapted to available systems.
	Allows patient to be handled in physical space	Allows personnel to communicate across physical space. .	Allows staff to communicate across physical space.	Constrains possible software maintenance via network.	-	-	AI receives input from network for navigation, operation, etc.	network needs to be able to enable emergency stop.	-	Technology in physcial space is controlled by software .	Software on network controls cooperating systems.	**Software environment**

Table 3.: N-squared diagram

3.5. Fishbone analysis of major faults

The vast possibilities of failure modes lead to the need of defining main failure modes. Using a fishbone diagram, an overview is given of the major fault modes that can occur, see [Table 4]. These are defined over the six principal views of the safety cube. Faults at the different levels are intrinsic to that specific level. Everything directly surrounding the AI is critical in that sense, especially the hardware. Some examples: sensors can fail, processors can fail, actuators can fail and the communication between those can fail. This is summarized as electrical and mechanical failure. Communication is not only important within the system itself, but also on the three principal levels of integration. Furthermore, it might deem very difficult to train people (especially patients) to work with AI cobots.

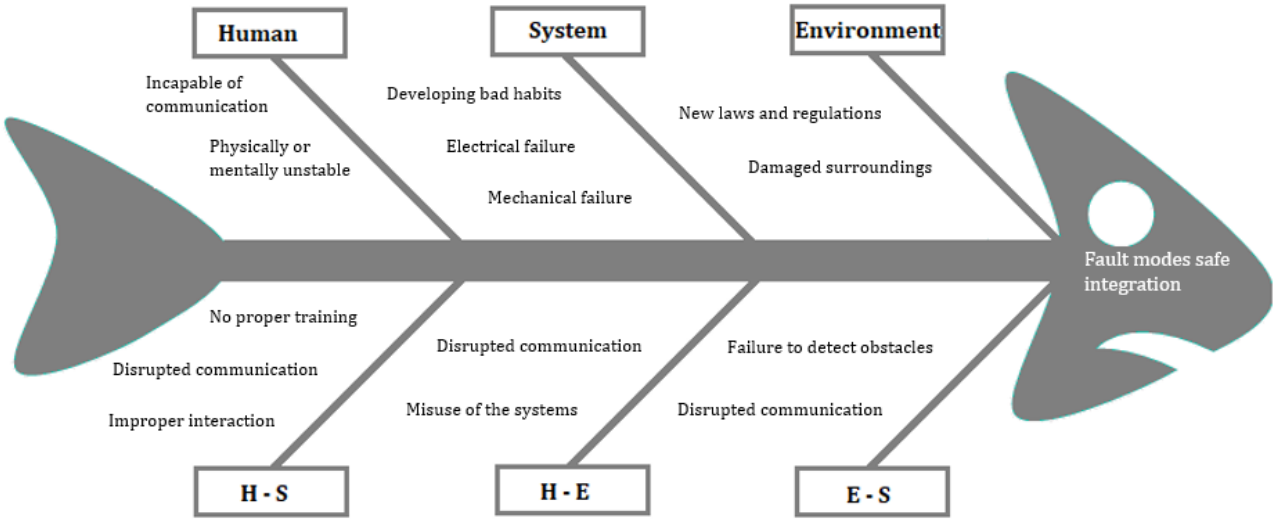

Table 4.: Fishbone diagram fault modes

3.6. Design constraints

The definition of a single design constraint (with inherent safety indicators) can prevent multiple fault modes from occurring or enables humans or the system to act accordingly. Some human intrinsic faults are impossible to address. Humans that are physically or mentally impaired or unstable might be helped by medication and technology, but in order to prevent health hazards by means of the cobot, it must learn extensively how to deal with such persons. AI technology in its current state is not yet, the law of acceleration predicts that that might be within a decade: *"My own predictions in terms of what technology will be able to do is that within less than 20 years, by 2029, computers will be at human levels, and by that I mean at the level of our emotional intelligence." [4]* This notion leads to the suggestion that more extensive laws and regulation on AI must be defined.

Regarding the failures within the mechanical and electrical domain there are rules and regulations that should be followed to make a safe design possible. Close attention to design of electrical circuits and mechanical parts, using well chosen safety factors is vital here. But also, the cobot itself is a weakness. It can develop bad habits just the same as humans. Supervised learning and teaching, just like children is therefore principal. As AI is not yet capable of complete human behaviour, dealing with environmental changes will prove difficult.

Patients with the above mentioned challenges may not react well to intelligent cobots or not even be able to learn how to deal with them. Proper training for both the patients and the cobot is thus key, especially with regards to communication and interfacing.

Humans also can misuse surrounding systems or might not be able to communicate with the systems, because of mechanical or electrical faults. This can have direct influence on the performance of the cobot, as it relies on data via sensors and network. Training for humans and inherent safe design is a solution here.

This also poses the problem that AI may have within the surroundings. Disrupted communication with the environment can cause the AI to not act as desired or expected. Again, inherent safe design is the principal solution. Lastly, as AI becomes more like humans (who have to obey laws), AI should also obey these laws or extended laws are necessary to guide these developments.

Judging these constraints, the following safety indicators emerge:

- Control of communication and intrinsic design of both the cobot, and environmental and cooperating systems. (Leading)
- Logging of emergency uses, failures and complaints. (Lagging)
- Proper training and education for cobot and humans. (Leading)

4. Discussion

The design constraints show once more that safe design with careful examination of laws, regulations and standards is important. The proposed attention on mechanical and electrical safety will lead to extra costs, but in this context, that is necessary. The latter constraints can mostly be addressed by following either the machinery directive or the low voltage directive. The derived constraints and indicators can be applied in other regions, this is also recommended by the authors. But keep in mind that this paper only gives a short overview of all the possible fault modes and that in the design process a full risk analysis according to a method as mentioned in [§3.4] or comparable is required.

5. Conclusion

In conclusion one can say that safe integration of AI enjoying cobots in healthcare technology is possible through careful design according to the defined design constraints, from which the following safety indicators emerge:

- Control of communication and intrinsic design of both the cobot, and environmental and cooperating systems. (Leading)
- Logging of emergency uses, failures and complaints. (Lagging)
- Proper training and education for cobot and humans. (Leading)

That is, an AI must be able to function on the same level as humans. The most important stakeholders are the patients, directly involved staff and personnel and the AI itself. These all expect safe and seamless integration and operation. Possible fault modes can occur on any of the faces of the safety cube and occur mainly in electrical/mechanical design, interfacing and communication. Current standards stay applicable, but additional standards are recommended. Interaction can remain safe through careful design and defined safety indicators.

References

[1] *2012 Department of Defense Standard Practice System Safety*, D. o. Defense, 2012.

[2] M. Rajabalinejad, "Safe integration for system of systems: The safety cube theory," 2019: Institute of Electrical and Electronics Engineers Inc., pp. 323-327, doi: 10.1109/SYSOSE.2019.8753867. [Online]. Available: https://ieeexplore.ieee.org/document/8753867/

[3] NASA, "Failure modes, effects and ciritcality analysis (FMECA)," 1999. [Online]. Available: https://kscddms.ksc.nasa.gov/Reliability/Preferred_practices.html.

[4] R. Kurzweil, "Interview with Ray Kurzweil. Interview by Vicki Glaser," *Rejuvenation research,* Article vol. 14, no. 5, pp. 567-572, 2011, doi: 10.1089/rej.2011.1278.