ICRES 2019: International Conference on Robot Ethics
and Standards, London, UK, 29-30 July 2019.
https://doi.org/10.13180/icres.2019.29-30.07.015

EXPLAINABLE AI, MODEL RECONCILIATION AND SYSTEM-LEVEL ANALYSIS FOR SAFE-CONTROL OF MEDICAL DEVICES

IOANNIS GEORGILAS
Department of Mechanical Engineering, University of Bath, Claverton Down
Bath, BA2 7AY, United Kingdom
E-mail: i.georgilas@bath.ac.uk
www.bath.ac.uk

There is an increase in the demand to healthcare systems to provide support for patients and give them a good quality of life. Given the limited human resources such support can be provided with specialised devices that can adapt to the needs of the patients. At the same time the number of prospective novel medical devices that are using AI is increasing every year. Only few will reach patients because of the difficulty to certify black-box systems. In this work we are proposing a method in which humans will work collaboratively with the AI system, building trust and collectively ensure the safe operation of the device. This method draws from the domain of Explainable Artificial Intelligence (XAI), model reconciliation, and System-Theoretical Process Analysis (STPA) to establish a transparent interaction and control regime. In this work the outline of the proposed system is given and how the different component will work and deliver the desired outcome. The ethical issues are also discussed and how the framework can sit within the existing regulatory setting as well as how the changes in standards for medical device certification and intelligent system will evolve.

1. Background

The number of patients that are facing significant challenges with regards to their ability to live independently with a good quality of life increases every year worldwide. These patients can benefit by specialised support which unfortunately cannot be provided because the necessary staff or resources are not available via strained healthcare systems. At the same time there is a growing number of medical devices that can assist those patients. Most of these devices are using intelligent methods based on Artificial Intelligence (AI) to predict and adapt to the needs of the patients. Unfortunately, this adaptability makes their certification very difficult and only a small number reach the market and often with their capabilities reduced.

The main reason for the limited adoption of AI in healthcare, and only for prediction of incidence occurrence and for diagnostic purposes off-line, is the lack of trust. This is because the latest and most powerful methods generate results in a black-box, difficult to understand way, and users and certification bodies cannot predict their operation and if and when an AI system will fail. The healthcare device domain is driven by safety and risk prediction to ensure that no harm will happen to the users. As a result, safety is assessed by the risk of failure of a system and in order for this to be evaluated exhaustive testing of all potential operational modes is the preferred method of safety certification. Because of the evolving nature of AI these operation modes are not known a priori and actually will change during use.

Despite these issues the use of AI will greatly benefit the healthcare domain and major actors in the certification and approval of medical devices started moving position in order to tackle them. Specifically the Food and Drugs Administration (FDA) in the USA is currently evaluating their procedures [1] with regards to AI-enabled medical devices in order to reimagine a regulatory approach for these devices. At the same time in the UK the National Health Service (NHS) and the department of Health and Social Care have published a code of conduct for data-driven technologies, trying to capture some of the issues associated with the use of AI in healthcare applications [2].

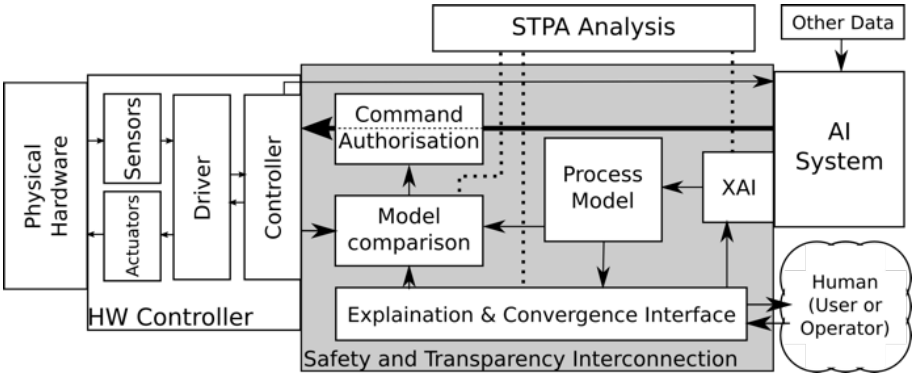

Figure 1. Diagram of the proposed system. With grey the new interconnection layer is noted.

These developments demonstrate that AI will need to be used in medical devices to actually control them. These enhanced systems will have faster reaction to altering conditions and most importantly adapt to better match user needs. For this to be realised it is crucial for a new method to ensure that AI-controlled medical devices can be trusted to remain safe. This work is a first step towards such a method which proposes that the human user/operator work collaboratively with the AI system, building trust and collectively ensure the safe operation of the device.

2. Proposed framework

Safety has been driving the development of new control methods for robotic systems for the past ten years [3] and there are existing safety standards like ISO10218 and ISO13482 to reinforce this. A key factor for improved safety is the ability of robots to explain their decision process in a transparent manner [4]. In the AI domain the use of explanations to improve trust is being studied with explainable AI (XAI) [5] and the overall desire to provide an explanation for black-box models [6]. A roadmap on how to use XAI in medical systems is also explored in [7] for the next generation of applications.

In this work we are proposing to expand XAI with the concept of model reconciliation [8] enabling the human and the AI system to collaboratively ensure safe operation. In order to facilitate this collaboration a common safe baseline will need to be established and this is achieved by analysing the safe operation of the system using a new safety assessment method, System Theoretic Process Analysis (STPA) that can be applied in complex systems [9] that also drives the design specification and process [10] .

In Figure 1 the proposed system is presented as a block diagram. On the left-hand side there is the physical hardware and its controller. Since this approach is aimed at the control of physical devices the presence of actuators and sensors is assumed as well as the relevant electronic drivers and low-level controller. On the right-hand side is the AI system predicting and adapting based on information from the controller (e.g. sensor data and status information) as well as other medical relevant data (e.g. medical history, clinical observations etc.) while the option to directly “observe” the user/patient (e.g. video input) is possible.

In the middle is the interconnection interface that will enable the collaboration with the human and ensure the safe operation. The first step will be performed in the XAI module to enable the generation of a process model that then can be compared with the process model of the human user/operator. The explanation and convergence process will take place in the interface of the interconnection. The function of the interface will be to explain the intention of the AI, compare it with the opinion of the human and if differing a question-answer process will

enable the convergence. If it is concluded that the models are agreeing, the commands issued by the AI module will be allowed to be executed by the hardware controller.

The parameters for the XAI module, comparison module, and interface will be derived based on an STPA analysis. The analysis will take part first, it will establish a process context parameter space as well as unsafe control actions within this space. Based on these actions and context information the generation of the models, their comparison and convergence will be facilitated. This way all scenarios will be covered, for example if both AI and human models are agreeing and within the safe action domain then the operation will be allowed. If they are agreeing but are outside the safe action domain then a re-evaluation and new convergence will be required. In the case that the respective models differ then first a convergence process will take place prior to establishing if the result is within the safe action domain.

3. Discussion

The proposed framework is aiming to create synergies between existing systems and methodologies. The crucial element of the proposed approach is the operation of the XAI module. This module is presented as a separate entity, outside of the AI core functionality because in this way it can incorporate a set of function. Primarily, and in line with existing use, it is ensuring that the decision process of the AI is captured and articulated in a comparable manner. It cannot by itself achieve the model reconciliation functionality. This operation must be conducted in a different process/module since it requires a different set of parameters and this is the function of the explanation and convergence module. For example, the latter can incorporate cognitive models of human decision making that can capture the intentions of the human operator while the XAI unit will need to emulate such models to "translate" the AI process.

The key function of the explanation and convergence module is to ensure a good communication between the AI system and the human user. Based on this communication the necessary trust will be developed and the proposed collaborative achievement of safety will be realised. The development of trust is not a linear operation and multiple steps need to be taken to facilitate this. It is envisioned that familiarisation with the system will be part of the framework. This means that beyond the technical development serious consideration must be given to the training regime and how such intelligent medical systems will be introduced to clinical environment. In its most simple form, a two-stage introduction will be needed, in the first phase the user will take part in training and simulation exercises where thy will learn the expected behaviour of the system and will acquire knowledge to assess the AI decisions as well as practice their own decision-making skills. This will enable them on the second stage, in the field, to quickly recognise appropriate decisions and will already have started establishing trust with the system.

Based on this close interaction with human users and the fact that the proposed framework will allow the operation of AI systems in close proximity to/and on humans a number of ethical considerations need to be taken into account and addressed. One of the key issues, as with most autonomous agents, is the degree of the AI system can take a decision that can have a consequence to a human. Although a complete answer cannot be given in this work and an extensive analysis is required, the structure of the system will enable the human operator to balance the AI's decision process. It can also be argued that the AI will have a similar effect to the human operator's decisions and the proposed system is envisioned to enable this balancing effect to ensure an ethical operation. Moreover, the fact that this system is designed to maximise the trust between the human and the AI system there is an argument about the ethical consequences of this. It must be noted though that the intention of generating trust is not in order

to create a false sense of security to the human. This might lead to a lapse of attention and thus to unsafe situations. It is intended to ensure that the human understands the system's operation and can trust that its outcomes will be as anticipated so collectively take decisions to maximise the benefits for the patients.

Another key issue is how the proposed framework sits within the context of certification of medical devices and the standards that dictate them as they stand now. It must be noted again that there is already a shift by the key certification bodies towards an acknowledgement that AI will be part of the healthcare profession in the next couple of years [1], [2]. This will likely lead to the adoption of a new set of standards to govern this process. These new standards most likely will build upon existing [11] and developing standards and guidelines [12]. The proposed framework is aligning well with the first standard (BS 8611:2016) since it aims to enable the learning AI system to maintain its safety by ensuring that the changes are easily identifiable individually. Since the framework will be developed in parallel to these new landscape it is possible to ensure and be able to demonstrate its accountability and thus facilitate the adoption of such systems in terms of the insurance risks and costs.

4. Next Steps

This work is aiming to provide a method with which new intelligent medical devices can use the power of AI to improve user/patient life. Here we have presented a brief outline of the key elements of the proposed framework and discussed some of the issues that might arise and will affect the feasibility and success of the proposal. The method presented is based on XAI, model reconciliation, and system-based safety analysis for allowing AI and human to collaboratively ensure safety by building trust of the human operator to the decision process of the AI.

The next steps to evolve this method is to investigate the STPA baseline process, assess different XAI approaches, and convergence methods with the aim to validate the system on real hardware. This will demonstrate the universality of the proposition and allow the field of medical device development to move forward, embrace novel techniques and methods.

References

[1] Food and Drug Administration, "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) -," 2019.

[2] National Health Service, "Code of conduct for data-driven health and care technology," 2019.

[3] R. Woodman, A. F. T. Winfield, C. Harper, and M. Fraser, "Building safer robots: Safety driven control," *Int. J. Rob. Res.*, vol. 31, no. 13, pp. 1603–1626, Nov. 2012.

[4] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, "Explainable Agents and Robots: Results from a Systematic Literature Review," *Proc. 18th Int. Conf. Auton. Agents MultiAgent Syst.*, pp. 1078–1088, 2019.

[5] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[6] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, Aug. 2018.

[7] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?," *arXiv Prepr.*, no. arXiv:1712.09923, 2017.

[8] T. Chakraborti, S. Sreedharan, S. Grover, and S. Kambhampati, "Plan Explanations as Model Reconciliation," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2019, pp. 258–266.

[9] I. Georgilas, G. Dagnino, and S. Dogramadzi, "Safe Human–Robot Interaction in

Medical Robotics: A case study on Robotic Fracture Surgery System," *J. Med. Robot. Res.*, vol. 02, no. 03, p. 1740008, Sep. 2017.
[10] I. Georgilas, G. Dagnino, P. Tarassoli, R. Atkins, and S. Dogramadzi, "Robot-Assisted Fracture Surgery: Surgical Requirements and System Design," *Ann. Biomed. Eng.*, vol. 46, no. 10, pp. 1637–1649, 2018.
[11] British Standards Institution., *Robots and robotic devices - Guide to the ethical design and application of robots and robotic systems BS 8611*. BSI, 2016.
[12] IEEE, "Ethically Aligned Design: A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems, Version 1.," 2016.