Design for Interpretability: Meeting the Certification Challenge for Surgical Robots*

Maria-Camilla Fiazza[1] and Paolo Fiorini[1]

Abstract— **This paper presents a perspective on some issues related to safety in the context of autonomous surgical robots. To meet the challenge of safety certification and bring about acceptance of the technology by the public, we propose principles for a design paradigm that goes in the direction of safety by construction: design with certification in mind, clearly distinguish the notion of safety from that of responsibility, view the human component as scaffolding in the progressive transfer of decision-making to the machine, preserve interpretability by renouncing black-box approaches, leverage interpretability to assign responsibility, and take corrective action only when the semantic of the human-machine interface is violated.**

I. Introduction

The field of medical robotics is currently engaged in a revolutionary push toward autonomy. The technology is intended to allow a physician to delegate progressively larger components of the surgical workflow to a machine. Eventually, the technology will result in a robot capable of executing an entire procedure autonomously, possibly even operating as part of a surgical team in the operating room.

This ambitious direction of research presents a set of challenges that go beyond the technical. In particular, acceptance and adoption of the technology depends critically on the ability to show that the robot is *safe* and that its deployment leads to performance at least not inferior to what can be achieved without it. The process of certification is intended to offer such guarantees to the public.

The main unresolved issue is the lack of a precise operational definition of safety in the context of surgical robots, and in particular in the context of *autonomous* surgical robots. A first key requirement of a suitable definition is that it would be applicable along all levels of autonomy, as classified in [1]. Secondly, it would include more then the mechanical and operational safety of the robot considered, for example, in [2], [3]. Because a robot at autonomy level 4 is in fact *practicing medicine*, safety ought to include the dimension of *patient safety*, which we have shown in [4] to be an emergent systemic property and to require an approach focused on system-wide dependencies and system integration.

A framework that allows for safety validation in such a wide variety of circumstances is a necessary step for surgical robots to overcome the hurdle of regulations and avoid rejection by the public as they reach market. Without both the reality and the acknowledgment that a surgical system is safe, deployment will not succeed. Regulatory requirements act as barriers to progress. Within a suitable certification framework, however, regulations become useful boundary specifications and can orient and guide forward technical development.

*This work has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No. 742671 "ARS")

[1]The authors are with the Department of Computer Science, University of Verona, Strada le Grazie, 15, 37134, Verona, Italy. {mariacamilla.fiazza, paolo.fiorini}@univr.it

II. Safety Paradigms

The standard process in use consists of first designing the system to meet its functional specification and then implementing safety safeguards within the paradigm of hazard assessment and risk mitigation. This is the logic employed, for example, by the recent ISO safety certification for assistive devices [5].

It may be necessary for system designers to also develop new evaluation methods specific to the domain and task. The process involves establishing quantitative safety bounds for all relevant features (pressure, temperature, etc.) of the physical human-robot interaction at each point of contact, for example as done in [6]. Validating safety then means showing that the machine operates within the bounds of the safety region thus established. This approach holds at the physical level and also at the logical level, whenever the safety region can be interpreted as the set of nodes (states) which are not connected to any nodes (states) known to represent an error condition.

These new methods, initially pertaining only to the device under development, can later be generalized to a class of devices with the same purpose or working mechanism. The resulting validation-and-testing toolbox eventually finds its way into regulations and standards, and ultimately becomes, years down the line, the basis for the work of certification officers.

As far as the systemic aspects are concerned—especially those intertwined with the fact that level 4 autonomous surgical robots are in fact practicing medicine—methods comprehensive enough to give rise to a standard have not yet been developed. It can be foreseen that there will be a substantial overlap with the way surgeons are licensed and certified by a medical board.

Designing systems with safety in mind depends critically on what is understood as a possible source of error. Each source of error is then tracked with self-monitoring architectures that detect deviations from expected functionality and are capable of implementing recovery actions. This approach has been pursued at all levels of abstraction, from hardware components [7] to high-level situation-awareness [8]. The

same supervisory role is played by the human under whose responsibility assistive surgical systems currently operate.

What does it mean to demonstrate safety in robots that by design do not operate in a controlled environment, do not rely on repetition, cannot constrain the range of human behaviors, and whose tasks are entirely context-dependent? In these circumstances, the systemic aspect of safety dominates and the key sources of error have to do with system integration and the logical properties of the interaction. Existing safety standards, on the other hand, mainly focus on a lower level of abstraction, that of the physical properties of the interaction.

III. Responsible Humans, Safe Systems?

The problem of safety in machines that operate semi-autonomously has been conflated with the issue of responsibility, probably because assumption of responsibility is a cornerstone element of legal thinking. Our approach clearly distinguishes the two concepts of safety and responsibility.

In the current regulatory environment, decisions made by the physician are by default acceptable, whereas the decisions made by the machine need to be proven safe. The disparity in the burden of proof is enormous: when the decision is entrusted to the machine, we are required not only to have the machine make the correct decision, but to *prove* that it will do so within a certain margin of risk. Methods to systematically evaluate the output are not available, at times the acceptable error is several orders of magnitude lower than what is afforded to humans, and therefore the unintended consequence of current regulations is to incentivize "offsetting to the human" in circumstances in which it is not a natural choice. As a result, the design problem itself is changed in artificial ways, taking the focus away from the development of needed enabling technology.

Often the human does not have access to better clinical information than the machine does and cannot double-check in meaningful ways. Users end up trusting the machine output that they are meant to oversee.

We interpret the human component in systems at autonomy levels 0-4 as needed scaffolding in the process of development and deployment of a level 5 device. The human's role is not to take responsibility for operation, but to:

1) fill in wherever the robot is not yet up to par with human performance, until the robot reaches that standard and can implement its own control,
2) serve as an embodied high-level monitoring architecture, and
3) provide a fail-over, an alternative mode of operation (e.g. laparoscopic or open surgery) that is well understood and that one can revert to in case of critical system failure.

Whereas the first is a functional role, the supervisory and fail-over elements are intrinsically safety features.

Rather than naively assuming that effective human supervision just happens, we hold that this supervisory role needs to be carefully designed, especially at higher levels of autonomy. Monitoring processes are tasked with continuously comparing expected output to actual output, for the entire duration of the procedure. In the case of surgery, procedures can last hours. Humans cannot sustain the level of attentiveness required of a monitoring process for such extended amounts of time, unless they are actively interacting with the monitored process. In surgical robots at autonomy level 3, the physician merely approves the robot's proposed plan of action, or chooses from a set of proposed alternatives, leaving the execution entirely in the robot's hands. At level 4 the human's contribution is purely supervisory. In both cases, for the greatest majority of the time the human is passively observing and runs the risk of incurring in attention fatigue and habituation.

System design needs to take into account the critical constraint of limited attention, and ask humans to supervise in the way they can do it best: discontinuously, paying attention at point of interaction in response to a relational prompt. Instead of relying on the hope of continuous human supervision, autonomous surgical robots realistically need adequate self-monitoring architectures that communicate through interrupts by raising flags and directing attention on specific situations. This also requires the machine to be able to notice that something is going awry. Knowing *what* is going awry requires very sophisticated thinking, but it is already possible for machines in the early stages of autonomy to notice *that* something is potentially going awry.

Semi-autonomous surgical systems must manage the flow of information and raise relevant safety-critical information in a way that primes the human to check using his/her own native perception and thought processes. When prompted and alerted to contextual information to integrate, humans provide a valuable additional level of safety via redundancy through diversity.

An additional benefit of this self-monitoring and communication style is providing the evidentiary record needed to resolve a case in court, an audit log. It is impossible to substantiate a claim of harm without having access to precisely the data that would need to be monitored intra-operatively to ensure safe decision-making.

IV. At the Interface

The boundary between what is controlled by the human and what by the machine is in all cases the critical element. In level 0 systems the human is so deeply in the loop that the machine's decision-making is reduced to choosing how to best translate the user input into motion, for example by performing tremor suppression. As autonomy increases, the human is less deeply in the loop and the jurisdiction of the machine grows.

Learning to drive offers a good model: there is another set of controls in the same car and an instructor is ready to take over until the apprentice is sufficiently trained to earn a driver's license. We believe that joint control is an inescapable and fundamental element on the way to fully autonomous systems. Once the technology reaches level 5,

the human interacts with the robot as an "outside" (collaborative) agent. Conceptually, the interface has moved and now separates distinct agents, instead of distinct subsystems in a joint control architecture.

Even though the interface moves, its properties should not change. At level 5, where humans are interacting collaboratively with robots, the key property is *interpretability*. We propose that this property must hold at all locations, and that preserving interpretability is a fundamental design constraint for safety-critical robots at all levels of autonomy.

V. Because I Understand You

Monitoring architectures and hazard analysis address the *how* of safety; in the context of surgical robots, at all levels of autonomy, the *what* of safety is decision-making. We propose that a safety-driven design process for autonomous machines be viewed in terms of the controlled and progressive transfer of decision-making from human to machine and that the safety certification process be viewed as certifying the process of decision-making, made transparent to the user (the intervening physician or the medical team) though semantic interfaces.

The careful design of such interfaces, possibly specified in the form of assume-guarantee contracts, allows tracing errors and undesirable outcomes to a specific agent behavior. The issue of responsibility is addressed by separating the user's operational responsibility—to behave in accordance with the semantics of the interface—and the responsibility for the system, shared by the developers and certification officers.

Whenever a specific area of decision-making has been placed on the machine side of the interface, the user should not be expected to take responsibility for the machine's decision-making in that area or intervene to correct it, unless it violates the semantic of the interface. The user often does not have access to the data necessary to assess the quality of the choice and should not be burdened with an impossible task. The detectable, behavioral violation of a contract is the trigger for corrective action.

To guarantee that smooth functional interaction between user and machine can occur, the design process needs to be structured and constrained in very significant ways from the start: the critical need is to limit what is visible to the user to what can be accomplished with interpretable tools. The most significant limitation is in the use of artificial neural networks (ANNs), with explainable decision making in the flavor of "explainable AI." Two main classes of approaches exist, according to [9]:

1) creating an apparatus to interrogate a black box and derive some of its properties or some properties of its output (post-hoc interpretability, e.g. in [10])
2) encapsulating a white box and restricting ANNs to using only interpretable primitives to approximate general functions (model-based interpretability, e.g. in [11]).

We favor the second approach. There is room to use black-box non-interpretable components as part of the scaffolding in the development process.

A second design constraint is that context needs to be organized and integrated systematically and the same context ought to be accessible by all modules. Safety cannot be "added" at the end by patching possible issues in the robot/human interface; rather, it emerges from the logical environment in which the system is immersed during development, as a fruit of system-wide consistency. The practice of going through a checklist of known sources of adverse events is still necessary, but it is no longer the core locus of safety.

The effort to preserve interpretability at all levels of the system allows a natural implementation for a number of important safety features, which we contend form a foundation for certifying the more abstract aspects of safety.

A safety certification process of the type we envision includes verifying at least the following:

- the existence of a logically consistent hierarchy of default behaviors, unambiguously determined by the circumstances. It must be possible to know in advance and with certainty to which behavior the robot will default in any given error condition.
- Mechanisms to log the intra-operative data necessary to derive correct recovery actions for rare and unexpected (but foreseeable) events, precalculated recovery plans, and algorithms to adjust such recovery plans intra-operatively within an acceptable time lag.
- Model- and knowledge-based methods to independently check the conclusions reached by the robot while using data-driven methods.
- The design of comprehensive audit logs, to allow forensic reconstruction of the decisions made during a given surgical procedure and to confirm what was communicated to the human user and when.

Validating the decision-making module and rendering predictable what the machine will do in any given set of conditions is the lion's share of the work. The requirement that all relevant information be available at the point of decision completes the picture of safety certification: given sufficient information to choose a good course of action, the robot indeed does so.

VI. Conclusion

One cannot guarantee safety in a system in which it is not known how to precisely assign responsibility for violations detected (or anticipated) during development. Because a root cause can have a multiplicity of consequences, often related to each other in obscure ways, the inability to trace the error to its decision-level origin introduces logical inconsistencies at the system level. They, in turn, can originate hard-to-see systemic errors and render the system unsafe.

On the other hand, when decisional causality can be made clear, responsibilities can be identified and corrective actions taken, through redesign or even during operation. In these circumstances, corrective actions are actual solutions, and not merely compensations for *some* of the effects introduced by an error upstream. This can be achieved by preserving

interpretability at all levels of the system, and by designing the flow of information between robot and user already within the paradigm of assume/guarantee contracts.

Certifying decision-making is a very big challenge. It will require innovation in key areas, because decision-making depends critically on both knowledge representation and context-awareness. Many things come together at this nexus, many distinct problems that require a unified approach because they are fundamentally aspects of the same problem: integrating ever-expanding knowledge from evidence-based medicine, developing autonomous surgical robots, reconciling data-driven and model-driven approaches, making decision-making explainable.

REFERENCES

[1] G.-Z. Yang et al., "Medical robotics—Regulatory, ethical, and legal considerations for increasing levels of autonomy," *Science Robotics*, vol. 2, no. 4, March 2017.

[2] *Robots and robotic devices – Safety requirements for industrial robots – Part 2: Robot systems and integration*, ISO 10218-2:2011, 2011.

[3] *Medical electrical equipment – Part 2-77: Particular requirements for the basic safety and essential performance of robotically assisted surgical equipment*, IEC 80601-2-77:2019, 2019.

[4] L. Grespan, P. Fiorini, and G. Colucci, "Patient Safety in Robotic Surgery," in *The Route to Patient Safety in Robotic Surgery*. Springer International Publishing, 2019, ch.2, pp. 7-23.

[5] *Robots and robotic devices – Safety requirements for personal care robots*, ISO 13482:2014, 2014.

[6] K. Homma, K. Fujiwara, I. Kajitani, and T. Ogure, "Safety Evaluation Technologies for Defecation Assistance Devices," in *IEEE International Conf. Intelligence and Safety for Robotics (ISR)*, Aug. 2018, pp. 446451.

[7] M. Li, Y. Ma, Z. Yin, M. Lian, and C. Wang, "A Structure Design of Safety PLC with Heterogeneous Redundant Dual-Processor," in *Proc. IEEE International Conf. Intelligence and Safety for Robotics (ISR)*, Aug. 2018, pp. 590594.

[8] M. Ginesi, D. Meli, A. Roberti, N. Sansonetto, and P. Fiorini, "Autonomous task planning and situation awareness in robotic surgery," in *Proc. IEEE International Conference on Intelligent Robots and Systems (IROS)*, Oct 2020.

[9] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. "Definitions, methods, and applications in interpretable machine learning," in *Proceedings of the National Academy of Sciences*, 2019, 116(44):2207122080.

[10] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri and F. Turini, "Factual and Counterfactual Explanations for Black Box Decision Making," in *IEEE Intelligent Systems*, vol. 34, no. 6, pp. 14-23, Nov.-Dec. 2019.

[11] Y. Yang, Z. Zheng, and W. E, "Interpretable Neural Networks for Panel Data Analysis in Economics," arXiv:2010.05311v3 [econ.EM], Nov 2020.