Levels of Autonomy & Safety Assurance for AI-based Clinical Decision Systems

Paul Festor[1,2], Ibrahim Habli[3,4], Yan Jia[3,4], Anthony Gordon[5], A. Aldo Faisal[1,2,6], and Matthieu Komorowski[1,5,6]

[1] UKRI Centre for Doctoral Training in AI for Healthcare, Imperial College London, UK
[2] Dept. of Computing, Dept. of Computing, Imperial College London, UK
[3] Assuring Autonomy Intl. Programme, University of York, UK
[4] Dept. of Computer Science, University of York, UK
[5] Dept. of Surgery & Cancer, Imperial College London, UK
[6] Dept. of Bioengineering, Dept. of Computing, Imperial College London, UK
[7] Corresponding author: m.komoroswki14@imperial.ac.uk

Abstract. Levels of Autonomy are an important guide to structure our thinking of capability, expectation and safety in autonomous systems. Here we focus on autonomy in the context of digital healthcare, where autonomy maps out differently to e.g. self-driving cars. Specifically we focus here on mapping levels of autonomy to clinical decision support systems and consider how these levels relate to safety assurance. We then explore the differences in the generation of safety evidence that exist between medical applications based on supervised learning (often used for prediction tasks such as in diagnosis and monitoring) and reinforcement learning (which we recently established as a way for AI-guided medical intervention). These latter systems have the potential to intervene on patients and should therefore be regarded as autonomous systems.

Keywords: Digital Healthcare · Autonomy · AI Safety

1 Introduction

Scales of increasing levels of autonomy have been proposed for AI applications in the healthcare domain [1, 14]. This raises a concern, especially for safety engineers, about how to define safety requirements for such autonomous systems. There is general principle that higher levels of complexity and autonomy imply more stringent safety requirements [2].

Supervised learning forms the basis of most AI-based clinical applications that have been proposed in the literature for diagnostic and monitoring purposes, e.g. in the recognition of skin tumors from images [3]. In this paradigm, the purpose of the algorithm is to try to predict as accurately as possible a defined prediction, labelling or classifications tasks using labelled data as training material - effectively making a single decision to the nature of patient's health state. In contrast, in reinforcement learning (RL), a agent learns a decision strategy (a so called policy) in a sequential decision making process. The policy is

optimised so that it maximises some form of future expected total return [13]. Formally it is related to optimal control and model predictive control applications. RL differs from conventional prediction tasks used in most of the medical AI literature in that the model does not simply reproduce human behaviour, it attempts to improve and learn an optimal decision strategy from sub-optimal training examples acquired from humans. This is a highly appealing approach for many clinical scenarios with uncertainty, since the method would - in principle - be able to tease out the right decisions among a range of options selected by human doctors. We developed such an algorithm in previous research for the treatment of sepsis (severe infections with organ failure), which we called the "AI Clinician" system and is now being developed for prospective clinical evaluation [7]. As we discuss in the later sections, the performance and safety assessment of the output of RL algorithms is more difficult than conventional prediction tasks based on supervised learning.

In this article, our objectives are to define increasing levels of autonomy of AI systems in healthcare and discuss differences between supervised and RL based AI applications with regards to safety assurance, using the example of the AI Clinician for sepsis treatment.

In autonomous systems the nature of the action of the agent is an important concept to consider. In autonomous vehicles the agent has to steer, accelerate and brake the car, give signals etc. However, in clinical settings it is not necessarily essential that the system directly controls the intervention.

2 Levels of Autonomy of AI Applications in Healthcare

We define 5 increasing levels of autonomy, from zero (no AI involved at all) to four (fully autonomous AI), as shown in Table 1. The differences between levels essentially reside in how the AI system and human user interact to make a decision, and who bears the responsibility for the decision made. We illustrate those levels using two different scenarios: self-driving vehicles, for their ease of visualisation and understanding, and clinical decision support systems, which are at the core of our research interests and whose deployment represents a true challenge today.

Level 0 is used to serve as a reference: it designates a version of the setup in which no AI is involved. For self-driving cars, this means human drivers without any assistance, and for clinical decision support system, this is standard care. This level is the baseline for the other ones: the aim of introducing AI is to increase some aspects of performance of the system so the performance should be assessed by reference to the baseline. This level can also serve as a reference for measuring the safety of the system. This makes sense particularly when considering systems which operate in high-risk environments, such as self-driving cars and clinical decision support systems (CDSSs).

Level 1 is the first step towards AI autonomy. In level 1, the AI system is set up in its environment, it can produce outputs and these outputs can be seen by the human agent. The human has the choice to decide whether or not to

Table 1. Definition of the five increasing levels of AI autonomy

Level	Short definition	Illustration in self-driving cars	Illustration in an AI for drug administration	Who makes the final decision/burden of responsibility
0	No AI	No assistance	Standard care	Human
1	AI suggests decisions to human	GPS guidance system suggests direction	Clinicians can see the AI recommendation	Human
2	AI makes decisions, with permanent human oversight	Car following lanes on the motorway, driver has to keep hands on the wheel	AI changes the doses, with human doctors continually checking	Human
3	AI makes decisions, with no continuous human oversight but human backup available	Autonomous car with a human behind the wheel, asks for driver's help if the AI cannot deal with the situation	AI changes the doses, and can alert human users in case of high uncertainty. Continuous human oversight not required.	AI
4	AI makes decisions, with no human backup available	Autonomous self driving car with no driving cockpit	AI changes the dose with no human backup	AI

look at the AI's output, and make their own decision accordingly. In this level, the requirements on the AI system are quite low as its dysfunction should have minimal impact on whether appropriate decisions are made or not. In the case of self-driving cars, consider GPS systems where the driver has the option to follow GPS guidance or not, and there is always a way for drivers to find their way without GPS, e.g. following signs and maps. Similarly, a level 1 AI-based CDSS would present treatment recommendations to the clinical team; however, the team should still be able to function without the AI.

Level 2 pushes the AI's autonomy one step further by letting it act directly on the environment. However, on level 2, the AI system is continuously monitored by a human expert who can take the lead at any time. An example of such an AI for self-driving cars is lane-following assistance where the car can change its steering and speed to stay in a given lane and maintain a reasonable distance from other vehicles. In the context of Clinical Decision Support Systems (CDSSs), a level 2 autonomous system would issue treatment recommendations which would be administered to the patient either directly or by a human. A human expert is continuously reviewing the system's output.

Level 3 represents what most people would call true "autonomy". Here, the AI acts directly on the environment, but it is not continuously monitored by humans anymore. Instead, the AI may request human input when needed. Note here that, because the interactions between human and AI are not as frequent or regular as in level 2, there can be delays in the human reaction, but these delays should be within an acceptable range for the application. This level introduces a new important requirement; the AI has to be uncertainty-aware, and be able to recognise states in which its output might not be appropriate and human

input is required. In the case of lane-keeping for self-driving cars, such a system would be able to keep the car safely on the road in most of the situations, without continuous human supervision, but could ask for human help when the conditions prevent the AI from being confident in its decisions, e.g. at a complex junction. Similarly, in a drug administration scenario, the AI would normally be able to take care of drug administration with no human supervision, but acknowledge the states in which the decision is not clear and ask for the input of a human expert. A typical example in healthcare applications is represented by mechanical ventilators that autonomously adapt to patient characteristics to accelerate the weaning process, e.g. [9].

Level 4 stands at the top of the scale and represents the state of complete autonomy. An AI system of level 4 autonomy acts directly on its environment, and should be able to handle every situation within its defined scope of use. This level is unrealistic with the current state of the art in both self-driving cars and CDSSs for drug administration and is questionable if this would ever be desirable in the healthcare setting [14].

An essential aspect of our approach is that the level of autonomy of a system can only be assessed with respect to the environment in which it operates and the tasks it aims to address. Even though there is a correlation between risk and AI autonomy, the absolute amount of risk involved in acting in different environments can be very different (an AI failing to mow the lawn properly has very different consequences from failing to keep a car in line on the motorway).

3 Difference between Supervised and Reinforcement Learning Applications

CDSSs based on supervised learning in the field of computer vision or closed-loops are already in use in clinical settings [14]. For example, the FDA has approved systems that detect strokes in brain CT scans and automatically alert physicians [11]. In the operating room, prototypes of AI closed-loop systems are being developed, to control the level of sedation (measured with real-time electroencephalography, EEG) or blood pressure during surgery [5,6,8], e.g. the amount of desired concentration of anesthetic in the brain (targeted controlled infusion, [12]). In our classification, such systems would correspond to level 1 (stroke detection) or 2 (drug dosing system). In the first example, the AI merely supports human clinicians in their decisions and improves workflow and care coordination. In the second example, the system controls the delivery of a drug that is being given to a human; it has latitude to increase or decrease the amount of drug flowing into the system. This is equivalent to a clinician ordering a certain amount of drug to be delivered and another human trying to control the amount injected accordingly. Yet, the clinician specifies the amount of drug in a part of the body, and the system is under the continuous supervision of a human expert (an anesthesiologist) who may take over the control of the system at any point.

Autonomy, however arises when the AI system is not directly working to precisely-defined human specification. An important distinction must be made

between AI-based clinical systems based on supervised learning and reinforcement learning. Let us use the example of the AI Clinician system for sepsis resuscitation to illustrate this distinction [7]. Sepsis represents a global healthcare challenge, a leading cause of death and the most expensive condition treated in hospitals [15]. A cornerstone of the treatment of severe infections is the administration of intravenous fluids and vasopressors. However, there is much debate around the dosing of these treatments and what resuscitation targets should be used. Despite decades of research, resuscitation strategies in an individual patient remain mostly empirical. This is the clinical challenge that the AI Clinician attempts to address. The challenge of collecting safety evidence is much more complex for an CDSS based on RL (such as the AI Clinician) than for the medical applications based on supervised learning described above, for a number of reasons.

Firstly, there is no established "gold standard" for sepsis treatment [15]. Clinicians may have multiple objectives, which run in parallel and might be conflicting. For example, optimising blood pressure with large volumes of intravenous fluids may temporarily improve cardiac output whilst compromising organ perfusion and increasing the risk of renal failure at a later time point. In a conventional supervised learning setting, the equivalent task would be to train a model to replicate desirable human behaviours, which is in general more straightforward. In sepsis resuscitation, the desired effect of fluid resuscitation and vasopressors may not be clearly defined, which makes it difficult to use supervised learning.

Secondly, while the RL agent can in theory explore treatment strategies that have not been used in practice by clinicians, there is in reality limited opportunity with RL to learn the optimal policy using "on-policy learning" by trial-and-error, due to ethical and patient safety risks [4]. A major limitation is the lack of high fidelity human simulators that would enable safe exploration of various decisions without compromising safety. In many real-world applications of RL such as healthcare, the environment in which to train the model is not fully observable, which induces uncertainty about the state represented by the RL model. This is very different from simulation frameworks used in computer science research such as the Atari games [10], where the environment is fully observable at any time and provides all the information needed to make optimal decisions.

Thirdly, the effect of the decisions on outcomes represents a complex closed-loop with confounded causality. The effect of administering fluids and/or vasopressors is realised at multiple time horizons on multiple parameters. For example, the effect on cardiac output and blood pressure can be immediate, while the effect on the patient's kidney function can be delayed by a few hours or days, and the patient's final outcome can be weeks away but still influenced by a single early decision. This is reflected when defining the RL model reward, where researchers have to choose between immediate, intermediate or delayed rewards.

Finally, key areas of our current work focus on the development of a safety case for the AI Clinician system in which the safety evidence and its associated arguments vary with the intended level of autonomy of the system and the

readiness of the wider clinical environment for the deployment of this novel technology.

Acknowledgements PF was supported by PhD studentship of the UKRI Centre for Doctoral Training in AI for Healthcare (EP/S023283/1). AAF was supported by an UKRI Turing AI Fellowship (EP/V025449/1). All authors acknowledge support by the Assuring Autonomy International Programme (Lloyd's Register Foundation and the University of York; Project Reference 03/19/07).

References

1. Bitterman, Danielle S et. Mak, R.H.: Approaching autonomy in medical artificial intelligence. The Lancet Digital Health **2**(9), e447–e449 (2020)
2. Catherine Menon, C.e., McDermid, J.: Defence standard 00-56 issue 4: Towards evidence-based safety standards. In: Safety-Critical Systems: Problems, Process and Practice, pp. 223–243. Springer (2009)
3. Esteva, Andre et. Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. nature **542**(7639), 115–118 (2017)
4. Habli, Ibrahim et. Porter, Z.: Artificial intelligence in health care: accountability and safety. Bulletin of the World Health Organization **98**(4), 251 (2020)
5. Joosten, Alexandre et. Rinehart, J.: Feasibility of closed-loop titration of norepinephrine infusion in patients undergoing moderate-and high-risk surgery. British journal of anaesthesia **123**(4), 430–438 (2019)
6. Joosten, Alexandre et. Rinehart, J.: Automated closed-loop versus manually controlled norepinephrine infusion in patients undergoing intermediate-to high-risk abdominal surgery: a randomised controlled trial. British Journal of Anaesthesia **126**(1), 210–218 (2021)
7. Komorowski, Matthieu et. Faisal, A.A.: The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nature medicine **24**(11), 1716–1720 (2018)
8. Lowery, C., Faisal, A.A.: Towards efficient, personalized anesthesia using continuous reinforcement learning for propofol infusion control. In: 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER). pp. 1414–1417. IEEE (2013)
9. Ltd, D.M.U.: Dräger smartcare®/ ps – the automated weaning protocol (2019)
10. Mnih, Volodymyr et. Hassabis, D.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015)
11. Murray, Nick M et. Hui, F.K.: Artificial intelligence to diagnose ischemic stroke and identify large vessel occlusions: a systematic review. Journal of NeuroInterventional Surgery **12**(2), 156–164 (2020)
12. Struys, Michel MRF et. Rolly, G.: Comparison of plasma compartment versus two methods for effect compartment–controlled target-controlled infusion for propofol. The Journal of the American Society of Anesthesiologists **92**(2), 399–399 (2000)
13. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. MIT press (2018)
14. Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. Nature medicine **25**(1), 44–56 (2019)
15. Yealy, Donald M et. Self, W.H.: Early care of adults with suspected sepsis in the emergency department and out-of-hospital environment: A consensus-based task force report. Annals of Emergency Medicine (2021)