1st International Workshop on Evaluating Progress in Artificial Intelligence - EPAI 2020
In conjunction with the 24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Risk assessment of artificial intelligence in autonomous machines

Agnes Delaborde[1]

Abstract. We consider the issue of AI (Artificial Intelligence) evaluation in the light of risk management for the proper marketing of machinery. When a physical device is driven by AI, the question arises as to how to manage the risks associated with the autonomous functions. Indeed, the risk of physical damage to persons or property in the event of a "wrong" decision must be properly assessed. The study presents the advances needed to carry out relevant risk analyses for the deployment of safe intelligent machines, and offers good practice that leverages the knowledge in AI evaluation to optimize the safety of these machines.

1 Introduction

The AI (Artificial Intelligence) and robotics scientific communities are highly active, and public authorities have taken up the subject of autonomous decision-making efficiently. There is however still a strong need for manufacturers to receive instructions on how to put safe and compliant AI-driven machines on the market. Among other obligations, manufacturers are required to carry out an analysis of the risks linked to the operation of the machinery; the risk assessment forms part of the mandatory CE marking file and must lead to the implementation of risk reduction strategies. More than a question of performance, AI evaluation may therefore be an issue of legal liability for the manufacturer.

This study is in the field of AI-based decision-making for machines (which includes embedded AI). This is a top-down approach: "I want to produce a machine that performs a task". The way of performing this task autonomously can indeed be realized with an AI algorithm. Machinery industry need references for performing risk assessments of AI-driven modules embedded in their hardware. Identification and quantification of the risk associated with AI is a major concern of the European Commission, as noted for example in the White Paper on Artificial Intelligence [5] issued at the beginning of this year, which warns of the need to list and categorize risks, and to put in place regulatory responses to the criticality of these risks. However, there is currently no approved method for quantifying the risks associated with autonomy, nor are there reference error rates and testing methods for AI.

At present, regulations and associated standards relevant to the manufacturer only encourage them to make sure that the system is functional, in other words that it does not crash. We understand, however, that the complex decision making enabled by AI can also lead to functional behaviors that are yet "wrong": these behaviors need to be identified and quantified in order to determine the extent to which the system may present risks. Manufacturers can choose either to develop their own AI systems (many open-source libraries are available for that), or to buy off-the-shelf AI solutions. In either case, manufacturers should be warned of the proper method to integrate this autonomy (allowed by AI) in their traditional risk assessment. Here, AI experts should play their part and bring their contribution to the risk assessment of the whole machine. This means that placing safe intelligent machines on the market requires bridging the gap between AI and machinery: manufacturers should encourage dialogue between computer scientists and roboticists in the design of safe machines. But they need to know how.

This paper proposes some approaches that may allow for reasonable risk management. In order to shed some light on the subject, we will begin by explaining the method used to carry out this exploratory study. The paper will then present a quick overview of the major issues linked to risk assessment for AI then offer a reasonable method for dealing with risk assessment. In the absence of current references for properly dealing with the safety of AI in machinery, this approach can allow the manufacturer to design their device based on sound reasoning and good practice.

2 Method and objectives of the study

The French national laboratory for metrology and testing (LNE), author of this study, is a French state-owned laboratory charged with the coordination of the French metrology, and the pre-market testing and certification of industrial products. The Department for the evaluation of AI systems (`https://www.lne.fr/en/testing/evaluation-artificial-intelligence-systems`) specializes in the design of metrology-grade evaluations for AI, including the development of dedicated software for evaluation, physical and simulation testbeds and the organization of evaluation campaigns. Over the years, the department has performed more than 850 evaluations of Information and Communication Technologies systems, and has led more than 30 national and international evaluations campaigns. Due to the interdisciplinary nature of AI, our domain of application is broad: natural language processing systems, agricultural robotics, industrial robotics, robot companions, autonomous vehicle, etc. Our status of independent evaluator therefore leads us to carry out evaluations to the benefit of industrial and public entities (developers, integrators, end users, etc.). According to our interlocutor's needs, the verification and assessment may concern for example performance, safety, reliability, or quality. When safety is concerned – and safety of AI is now a major concern – we achieve this by relying, where possible, on

[1] LNE - Laboratoire national de métrologie et d'essais (*French national laboratory for metrology and testing*), France, agnes.delaborde@lne.fr

regulatory requirements and normative recommendations.

The study was initially engaged through an analysis of autonomous machines for agriculture, whose supposed dangerousness (related to the size of the device, their effectors, the deployment in open fields, etc.) generates major regulatory and societal concerns. This analysis implied an estimation of the adaptability of the current regulatory and normative framework to the autonomous functions of these machines; this framework states in particular that the first stage towards compliance is the risk assessment. Transversally to our various research projects and partnerships, we have gradually come to the observation that the stage of risk assessment presents a difficulty that is common to all application sectors, and that it could also prove useful to the evaluation. Indeed, whether it concerns purely software AI (natural language processing, etc.) or AI-driven modules in machines (computer vision, task planning, etc.), the questions are always the same for the evaluator: how thorough should the evaluation be? What level of performance is valid? Does the test protocol cover all possible system behaviours? What are the "important" behaviours to verify? Are there any "wrong" behaviours that need to be compulsorily verified? The list is not exhaustive. A first answer emerged however, which was common to all the evaluations: the risk analysis of the AI system should be the entry point for the development of the evaluation protocol.

Indeed, the strategy applied in risk management may provide a list of undesirable (or even "wrong") behaviours, which therefore may indicate the elements that need to be checked in the evaluation procedure. The impact (on humans, goods, other parts of the system, etc.) of these behaviours may also help in the definition of the weight of errors in the final evaluation scores. Systematic risk analysis is obviously not new in the field of software. We find it for example in the IEC 61508 standard for programmable electronic devices [9]. However, its application to AI evaluation is far from being systematic. The study presented here explores the extent to which risk assessment can become a must-have tool in the AI evaluator toolbox. One should understand that these trails are not yet proven through rigorous scientific experimentation: they are the fruit of experience, submitted to the community and intended to lead to other projects pushing the exploration.

3 Major issues of risk assessment for AI

3.1 An emerging domain

For about ten years now, AI has been used to facilitate risk analysis, in particular for processing large volumes of data in order to predict risks related to a specific professional expertise. One can cite, for example, the assessment of risks for complications in patients treated with chemotherapy [7], the analysis of the risk of ground water contamination [16] or the design of AI methods to deal with credit risk [3]. Safety concerns are pretty well covered in the field of autonomous vehicle, where many studies tackle the issue of risk assessments [15, 14]. Here again, the algorithms are expected to model or predict the level of risk presented by driving situations. But the reverse, i. e. performing the risk analysis of AI itself, in particular when embedded in machinery, is an emerging discipline.

Among the initiatives, we can cite the COVR European project [2], that aims at providing methods for the development of safe robots. This ongoing project provides grids and tools for the assessment of risks in collaborative robots. But the framework deals rather with the general behaviors of robots, without pointing specifically to the AI functions.

One may note that, on the whole, software aspects of the machines are under-explored in regulation and standards. Machines are expected to be compliant with the EC Directive on Machinery [6], for which public enquiries and national working groups have revealed possible shortcomings with regard to the capacity for autonomy allowed by AI. On the normative scene, we note that different standards deal more or less thoroughly with the verification of autonomous functions: for example in mining machinery [11], in highly automated agricultural machinery [12], or collaborative industrial robots [13]. These standards provide interesting and relevant approaches for the identification and testing of certain autonomous functions in their sector. There is, however, a need for an overarching standard setting out the general method for risk management in autonomous machine functions.

3.2 Risk identification

Traditional risk analysis rarely explores deeply the software aspects of machines. Indeed, the texts do not encourage systematic exploration of software performance, except when it comes to systems performing safety functions. The usual exploration is generally limited to performing a verification of functional performance, in which the point of verification would be to determine whether or not the system crashes. However, in the case of "intelligent" decision-making systems, it is also relevant to ensure that the "right" decision is made. System performance certainly has a strong impact on the quality of the system – this will be an economic criterion – but it is not difficult to find examples where under-performance can cause damage, which leads to the domains of safety and legal liability. Here is a trivial example that constitutes a danger to property: "The detection system has wrongly detected a weed, and starts the activation of the hoeing tool whereas it is a crop plant, resulting in its destruction". In this case, the seriousness of the damage is very low. But what if the system wrongly decides to activate the hoeing tool while a human being is under the machine?

A systematic identification of the dangers linked to decision-making is therefore necessary, and this exploration requires a specific expertise that may not be present among the manufacturer's resources. This possible poor identification of AI risks may be due either to the manufacturer's lack of awareness of the risks associated with AI under-performance, or because the AI module is purchased off-the-shelf and its characteristics are not fully controlled or known by the manufacturer, and/or due to the general lack of reference methods for AI verification and validation.

3.3 Criticality assessment

Additionally, there are no systematic methods for determining "how dangerous" a failure of an autonomous function driven by AI can be. This criticality, usually represented by a numerical value calculated on $Damage\ Severity \times Exposure\ Frequency$, therefore requires prior quantification of the probability of occurrence of the hazard, which means in this case the probability of a failure.

In the design of an autonomous device, if a functional AI module can make a decision that results in damage, then the risk criticality calculation rule may apply, and the exposure frequency can possibly

be interpreted as the error rate of the module. Error rates (and performance rates) are traditionally computed by the AI developers so as to demonstrate the system's efficiency and to guide the tuning of the algorithms. Metrics such as accuracy, F-measure are commonly used, as well as methods for test data sampling. Here again, however, the manufacturer who embeds AI can easily be stuck, since there are few to no official references for the computation of error rates.

In the context of AI for safety components, the system should be tested so as to demonstrate that the error rates are within the PL (Performance Level) intervals specified in the ISO 13849-1 [10] (unless more specific standards apply). This means that the probability of failure should not exceed a certain amount, if one wants to guarantee the safety of the behavior of the component. However, the applicability of this standard to AI algorithms has not been demonstrated. Various research projects currently underway at LNE tend to show that the expected error rates may not be achievable by AI systems, which would probably indicate that: a) either the recommended test methods and thresholds must be adapted to AI, b) or that it is strictly necessary to surround AI modules with safeguards (software or hardware barriers), which can be a serious impediment to the economic development of AI for safety of machinery.

4 Adapting risk assessment to the manufacturer

4.1 Which risk assessment method?

The approach adopted in this study for the identification of risks is based on FMECA (Failure Mode, Effects and Criticality Analysis). This analytic risk assessment method is based on the decomposition of any product in smaller parts (components or functions), which allows the identification of the failures that can happen to each of these parts. The objective here is not to explain the entire implementation of a FMECA, but to identify the points of the analysis that must be handled in a particular way because of the autonomy of the machines. This aspect will be explored in more details in the strategies that we offer in Section 5.

Choosing the FMECA approach is justified on the one hand by the fact it is well-known by manufacturers, which can guarantee easy implementation, and also on the availability of taxonomies of functionalities related to decision-making autonomy. There is no real consensus in the community on "one" taxonomy for autonomous functions, because the categorisation is carried out according to the needs of the designer of the taxonomy, in order to put forward one or another aspect: the type of algorithm, the purpose of the algorithm, or the type of inputs, etc. However, there is a lot of overlap between taxonomies, and the main thing is simply to follow good practice: the decomposition must be fine enough to identify the smallest element whose state may affect the safety of the whole system.

4.2 Two approaches to embedded AI

The approach to risk assessment, for a specific AI module, will be performed differently by the manufacturer depending whether the module is bought off-the-shelf, or if it is homemade.

If the AI module is bought, the manufacturer should: a) Check that the technical specifications of the element cover the intended use of the machine; b) Verify that error rates are available (with indications on the test method), and take these rates into account in the risk analysis; c) Verify that the element has a safety certificate of adequate level if it is integrated in a safety component.

If the AI module is homemade, the manufacturer should: a) Ensure that the design process complies with quality specifications. Pay particular attention to the fact that the software has been designed following a strict quality protocol; b) Perform tests (or a formal verification when possible) to determine the error rate of the module, and take this rate into account in the risk analysis (a high error rate, combined with a high level of damage severity, will require a drastic risk reduction).

5 Strategies for risk assessment of AI

5.1 Setting up of the risk analysis

Systematic analysis of the risks linked to the autonomy of the machines can be based on four steps (see Figure 1). The entry point is therefore to identify and list the modules of decision-making autonomy. For each module, it will be necessary to determine the causal chain (does the operation of this module have an impact on this or that function of the machine), to identify the failure modes associated with autonomy, and to estimate the probability of occurrence of a failure that could cause damage.

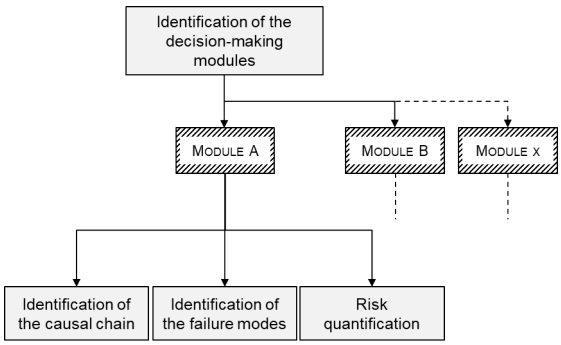

Figure 1. General approach for risk assessment of autonomous decision-making modules

5.2 Identification of the modules of decision-making autonomy

Let us start with a definition: a module of decision-making autonomy is a system (in the strict sense of the term) performing autonomous processing on input data in order to produce output behaviour(s). Note that the input data can be provided by a human (case of a man-machine interface).

Defining what is "autonomous processing" in a machine is not a trivial task. A sequence of mechanical actions can be performed without human intervention, yet it is not the type of functions that are concerned by the study. We will consider that an autonomous function necessarily implies a decision making of a software nature (not simply mechanical or electronic). Programmable logic circuits, considered alone, would therefore not constitute a module of decision-making autonomy, since they do not require software programming as such. The same applies to a sensor such as a Lidar, which simply helps to interpret the distance to an object.

We note that these elements (programmable logic circuit, Lidar scanner, etc.) are nevertheless elements of the decision module that must be verified, as they have a direct impact on the quality of the decision making. Indeed, they can be inputs for the decision-making process. Example: a Lidar sends an obstacle detection signal, from which the mobility management system takes the decision to stop. It is therefore essential to ensure that the Lidar has an acceptable rate of performance.

A taxonomy of functionalities is used to direct the search for the modules of decision-making autonomy. It should be noted that these functionalities are not exclusive: for example, a module enabling mobility does not mean that it does not perform a detection action for this purpose. Each module must then be decomposed in order to identify, at the end of the decompositions, all the modules present in the machine.

5.3 Identification of the causal chain

For each identified functionality, it is necessary to break down each module into three components: “input”, “processing”, and “output”. Processing is the decision function performed autonomously, regardless of the type of decision or the implementation mode: it can be an algorithm for classification, prediction, etc., implemented by a probabilistic decision tree, a set of expert rules in Boolean or fuzzy logic, a neural network, etc. Each input, and each output, can correspond to : another module (which it is necessary to decompose in turn); a physical component related to the machine’s capture (bumpers, scanners, encoders, etc.); a physical component linked to the machine controls (mobility effector, task effector, etc.) or to the Human-Machine Interface (inbound or outbound interaction with the user).

The causal chain between the individual modules and components must be represented, for example, in the form of a tree. This ensures that : 1) Each component has been identified, and that the analysis of its failures can be carried out; 2) Each system has been identified, and that the analysis of its failures can be carried out; 3) The impact of a failure of one component on another component is identified. Figure 2 shows an example of the relationship between two functionalities: weed identification and weeding task planning.

It is essential that the tree also includes the other components of the machine, even if they are not directly part of a decision-making autonomy module, in order to identify whether their malfunction can have a more or less direct impact on the module. As pictured in the example, note that detection (camera) and mobility (hoeing effector) are not counted as autonomous functionalities, because we do not locate autonomous decisions at their level, but they still appear in the tree.

We can also note that the “Classification algorithm” processing can be further decomposed, depending on the approach chosen by the developer. For example, the operation can start with a binarization algorithm to distinguish plants from the rest of the image (soil, etc.), then a machine learning algorithm will perform the identification of those parts recognized as “plants”. It is up to the developer of the algorithm to decompose the different modules of his system.

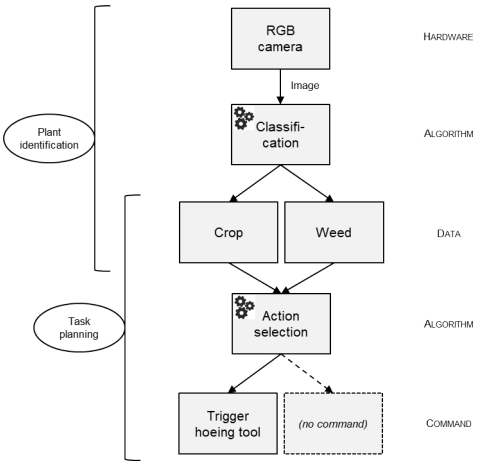

Figure 2. Input and output of two functionalities: plant identification and weeding task planning

5.4 Identification of failure modes

Once the tree has been built, each element must be inspected for possible failures. These failures represent the cases where the element does not perform the expected behaviour. All “classical” failure modes must be considered, such as unintentional start or stop, impossible start or stop, degraded operation. These aspects will not be discussed further here, since they do not relate exclusively to software. In addition, they are stated clearly in the EC Directive on Machinery, and seem to be part of the reasoning mechanisms well integrated by manufacturers. We recommend that these causes of failure should be specifically studied:

- **Software “crash”.** An error in the software programming leads to a malfunction: a critical shutdown or putting the software in an incorrect logical state. This may be related to the “scripting” of the software (the management of the transition from one state to another), a logic error, etc. These errors must normally be detected and corrected during development, notably by unit tests. It is up to the developer, or the QA manager, to show that these tests have been carried out.
- **Underperformance of the decision algorithm.** Probabilistic systems (identification, prediction, classification, etc.) rarely reach error rates of 0%. Complex non-probabilistic systems may also have some error rate that is difficult to correct. It is necessary to identify possible decision errors.
- **Misuse.** The human being is in the loop, either as an operator of the machine (remote control, collaborative work, troubleshooting action, etc.) or as a worker at the side of the machine. For each component, it must be identified whether the human being can cause a failure through reasonably foreseeable mistake or carelessness.
- **Difficult environmental conditions.** The environment in which the machine is deployed may impact the components. Environmental conditions are specific to the intended application of the machine and should be listed (fog, rain, mud, etc.). Indoor conditions shall be considered if this corresponds to the intended application (low light, narrow doorways, etc.).
- **The lack of explainability.** This point may be more difficult to apprehend, while possibly being ahead of future regulatory requirements for AI, as stated by the EC report about robustness and explainability for AI [4]. It is a question of verifying whether a lack

of understanding, on the part of the human operator, concerning the decision-making carried out by a "processing" component (a decision algorithm) could generate an undesired human reaction (effect of surprise, stress, overcompensation) which would then lead to a "risky use" of the machine. The risk reduction strategy would then consist, for example, in making the decision more "explainable", more "anticipable", and in improving ergonomics.

5.5 Risk quantification method

We present here an empirical study on different risk quantification methods proposed by the IEC 31010 standard on risk management [8], in order to estimate their adaptability to computation of error rates for AI.

For example, the Markovian approach allows analysing systems with dynamic behaviour, i. e. which can change states. This state-based analytical approach is based on the determination of the different states of the system, and on the probabilities of switching from one state to another. However, in the case of automatic learning for example, we note i) that the number of input parameters of the system quickly leads to a combinatory explosion, because the approach tests all combinations stochastically, (ii) that the identification of the parameters of greater weight (in order to limit the combinatory explosion in particular) is still a matter for research, for example in the field of research relating to the explainability of algorithms, and finally, (iii) that the possibility of controlling these parameters may be non-existent, for example in the case where the algorithm is off-the-shelf. The Markovian approach seems promising for the risk analysis of AI, but further research is needed to generalize its applicability.

The same applies to the Monte Carlo simulation risk analysis method, which aims to simulate system behaviour by injecting random input values over a large number of repetitions. The results thus provide ranges of values of possible behaviors. In the case of AI again, this method requires research, in particular to determine the nature of the input values.

A third method can consist of tests on representative samples. In this case, the determination of the test database is a challenge: it is a question of determining all the factors influencing the system's decision-making, determining the relevant configurations of the input data, and fixing the distribution of these configurations within a test database on a "human" scale, based in particular on the severity of the risks associated with a specific configuration. The notion of "human" scale refers in particular to the time required to acquire, prepare and qualify the data in the test database, which may not be achievable automatically. For the moment, there is no normative reference for the testing of software modules with AI; these tests are based on the evaluator's regulatory skills and knowledge, by consensus with the manufacturer.

In this document, the choice of the quantification method is not formally stated. Firstly, because the appropriate method of quantification may differ depending on the nature of the system, and also because the methods identified require practical expertise in statistics, computer science and data science. Quantification can be carried out internally, at the manufacturer's premises, for example by the system developer if they have the necessary skills and resources to carry out this analysis. Consideration may also be given to the recourse to third party expert evaluators.

5.6 Error rate computation

The error rate should be considered in regards with the damage severity. In this view, a strong level of criticality (as pictured in Figure 3) will require a stronger remedial strategy to minimize residual risk.

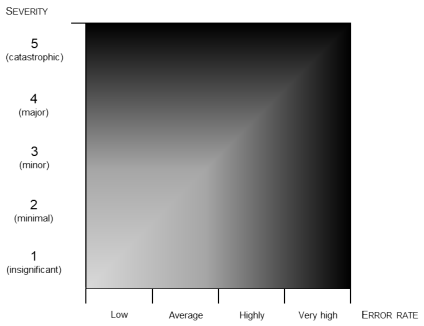

Figure 3. Criticality Matrix: relating the severity of damage to the error rate of the autonomous module of decision-making. The darker the area, the more critical the damage.

Voluntarily, the error rate presented in the Figure 3 is not quantified. Indeed, the value and interpretation of an error rate is dependent on first, the nature of the functionality: some technologies are particularly advanced, and it is common to observe very good performance scores. The quality of these performance scores must be estimated in relation to the state of the art, i. e. the scores obtained by other members of the community under the same conditions. This information is not always available, even though open competitions and test campaigns in AI and robotics are booming. Secondly, the estimation method also have an impact: the nature and number of test data, test method, metrics used, experimental conditions (laboratory, "real life", etc.) differ according to the type of algorithm considered.

The definition of an appropriate error rate should be based on the identification of the relevant indicators of under-performance. These indicators can be represented by types of errors such as :

- **False positive.** The system has an activation when it should not have. Example: The presence of an object was detected when there was nothing. Example of the concerned functionality: Detection.
- **False negatives.** The system did not present an activation when it should have. Example: An object was present but its presence was not detected. Example of affected functionality: Detection.
- **Misclassification errors.** The system returns one value when another was expected. Example: An object is confused with another one. Example of concerned functionality: Identification.

These errors, combined with the severity of the hazard, will be used to determine the metric for quantifying the error rate. For example, if the device is likely to run over an individual (high severity) in the event of misdetection, this means that the appropriate metric for risk assessment will need to be a function that correctly represents false negatives (a recall function is a good example). This process of metric selection is traditionally applied for the evaluation of functional performance, yet in a different spirit. For example, in the French robotics competition ROSE for agriculture [1], one indicator of performance is related to the

appropriate classification of crops and weeds, and the corresponding metric highlights the element of importance, which is, in this case, the economic efficiency of the machine (i. e. not destroy crops).

At the present time, in the absence of official thresholds or methods, we advise to interpret the error rates in an "expert" way. A developer experienced in his field will generally be able to identify the metrics that can describe the performance rate of the developed algorithms, and estimate whether the error rate is excessive or not. This analysis can also be performed by a third-party evaluator or an auditor.

6 CONCLUSION

We note that there are a number of "hollow points" in risk assessment for AI-driven autonomous functions for machinery. Most of them concern the requirements expecting a quantification, or a precise estimation and verification. Risk quantification is one of the major issues, since there are yet no reference benchmark methods for the estimation of AI performance.

However, it is possible to carry out "reasonable" risk analyses, to optimize the deployment of safe AI-driven physical devices. It seems essential that manufacturers should ensure that: a) They absolutely deepen the analysis of risks related to software aspects when AI is concerned, as described in this document; b) They encourage, within the company, a closer relationship between QA/safety experts and the developers of robotic and software platforms. If one of the expertise is not present, the call to a third party organization can allow an accompaniment on different points (normative and regulatory watch, AI and robotics evaluation, etc.); c) They consider the fact that a safety component must obtain a safety certification, and that this also concerns a component using AI; d) They are ready to provide software or hardware "safeguards" that can reduce the risks associated with any autonomous functionality whose error rate cannot be computed with sufficient reliability.

We notice in our regular interactions with stakeholders in the field of safety for AI, that while it is strictly necessary to continue regulatory and normative work, there is also a need to rationalize the fears generated by AI in machines. Indeed, formally representing all possible behaviours of an AI is not yet within our reach, which may lead to unexpected dangerous behaviours. Similarly, we still know little about the reactions of individuals in the presence of autonomous machines (curiosity, voluntary or involuntary misuse, etc.). However, this does not mean that human common sense cannot compensate for the unknown by adopting simple and effective risk reduction strategies, which is proven by our many interactions with manufacturers.

Offers from organisations providing support on these subjects are currently being developed, in parallel with changes in standards and regulations (changes in the Machinery Directive, standard IEC 61508 [9], standards currently under production by the ISO/IEC JTC 1/SC 42 commission on Artificial intelligence). We note that certain points are still at the research stage, such as the formal verification of AI algorithms or explainability, but that they are potentially in the process of becoming regulatory requirements. It therefore appears necessary to tie the link between the manufacturers of autonomous AI-driven machines and the AI community, in order to enable the deployment of safe machines.

ACKNOWLEDGEMENTS

As stated in this present paper, this exploration is based on many of our past and present research projects and partnerships. However, we would like to highlight the research projects that have allowed us to deepen this reflection in a very concrete way. This work was partly funded by the SOLROB project (2019-2020), in response to a call for tenders from the French Ministry of Agriculture via France Agrimer, led by the French association for agricultural robotics RobAgri. This work was also partly funded by the cascade grants ECAI and Blaxtair Safe, funded by the H2020 COVR (EU grant ID 779966, 2018-2021) on safety of collaborative robotics.

REFERENCES

[1] Guillaume Avrin, Agnes Delaborde, Olivier Galibert, and Daniel Boffety, 'Boosting agricultural scientific research and innovation through challenges: the rose challenge example', *3rd Rendez-Vous Techniques AXEMA, Axema, Paris Nord Villepinte, France*, **201**, 9, (2019).

[2] Jule Bessler, Leendert Schaake, Catherine Bidard, Jaap H Buurke, Aske EB Lassen, Kurt Nielsen, José Saenz, and Federico Vicentini, 'Covr–towards simplified evaluation and validation of collaborative robotics applications across a wide range of domains based on robot safety skills', in *International Symposium on Wearable Robotics*, pp. 123–126. Springer, (2018).

[3] Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock, 'Explainable ai in fintech risk management', *Frontiers in Artificial Intelligence*, **3**, 26, (2020).

[4] European Commission, 'Robustness and Explainability of Artificial Intelligence', Technical report, (2020).

[5] European Commission, 'White Paper on Artificial Intelligence: a European approach to excellence and trust (COM(2020) 65 Final)', Technical report, (02 2020).

[6] EU Machinery Directive, '42/EC of the European Parliament and the Council of 17 May 2006 on machinery, and amending Directive 95/16/EC (recast)', *Off. J. Eur. Union L*, **157**, 24–86, (2006).

[7] Patrizia Ferroni, Fabio Massimo Zanzotto, Noemi Scarpato, Silvia Riondino, Umberto Nanni, Mario Roselli, and Fiorella Guadagni, 'Risk assessment for venous thromboembolism in chemotherapy-treated ambulatory cancer patients: a machine learning approach', *Medical Decision Making*, **37**(2), 234–242, (2017).

[8] IEC 31010:2019, 'Risk management — Risk assessment techniques', *ISO/TC 262 Risk management*, (06 2019).

[9] IEC 61508:2010, 'Functional safety of electrical/electronic/programmable electronic safety-related systems', *IEC SC 65 A: Industrial-process measurement, control and automation - Systems aspects*, (4 2010).

[10] ISO 13849:2015, 'Safety of machinery — safety-related parts of control systems', *ISO/TC 199 Safety of machinery*, (12 2015).

[11] ISO 17757:2017, 'Earth-moving machinery and mining — Autonomous and semi-autonomous machine system safety', *ISO/TC 127/SC 2 Safety, ergonomics and general requirements*, (2017).

[12] ISO 18497:2018, 'Agricultural machinery and tractors — Safety of highly automated agricultural machines — Principles for design', *ISO/TC 23/SC 3 Safety and comfort*, (2018).

[13] ISO/TS 15066:2016, 'Robots and robotic devices — Collaborative robots', *ISO/TC 299 Robotics*, (2016).

[14] Kibeom Lee and Dongsuk Kum, 'Collision avoidance/mitigation system: Motion planning of autonomous vehicle via predictive occupancy map', *IEEE Access*, **7**, 52846–52857, (2019).

[15] Stéphanie Lefèvre, Dizan Vasquez, and Christian Laugier, 'A survey on motion prediction and risk assessment for intelligent vehicles', *ROBOMECH journal*, **1**(1), 1, (2014).

[16] Farzaneh Sajedi-Hosseini, Arash Malekian, Bahram Choubin, Omid Rahmati, Sabrina Cipullo, Frederic Coulon, and Biswajeet Pradhan, 'A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination', *Science of the total environment*, **644**, 954–962, (2018).