ESSENTIAL H&S REQUIREMENTS FOR INDUSTRIAL MACHINES EQUIPPED WITH MACHINE LEARNING

Date
14 September 2018

TNO innovation for life

Report for
Ministry of Social Affairs and Employment

ESSENTIAL H&S REQUIREMENTS FOR INDUSTRIAL MACHINES EQUIPPED WITH MACHINE LEARNING

Report for	Ministry of Social Affairs and Employment
Date	14 September 2018
Authors	Wouter Steijn, Liisa Janssens, Jan Harmen Kwantes, Dolf van der Beek & Anne Jansen (project manager)
Project number	060.31545
Report number	TNO 2018 R10499
Project name	MAPA Robotica 18.204.2-11

TNO contact	Anne Jansen
Telephone	+31-(0)8886 60991
Email	Anne.jansen@tno.nl

Contents

List of Abbreviations

AI	– Artificial intelligence
CE	– Conformité Européenne
EFTA	– European Free Trade Association
EU	– European Union
FIM	– Futuristic industrial machines
GDPR	– General Data Protection Regulation
AIM	– Advanced industrial machines
HCI	– Human-computer interaction
HLMI	– High-level Machine Intelligence
H&S	– Health and safety
ISO	– International Organization for Standardization
MD	– Machinery Directive
MiFID	– Markets in Financial Instruments Directive
SME	– Small and medium-sized enterprises
MRA	– Mutual Recognition agreement
OSHA	– Occupational Safety & Health Administration
R&D	– Research and Development
RL	– Reinforcement learning
SZW	– Ministry of Social Affairs and Employment
SML	– Supervised machine learning
TIM	– Traditional industrial machines
UML	– Unsupervised machine learning

1 Introduction

The robotics industry is developing rapidly in a wide range of sectors, from healthcare to manufacturing. Over the past fifty years, industrial machines have become faster and more accurate. They have also become more mobile. These machines now have more degrees of movement than ever before, and – thanks to improved sensors – they are becoming ever better at interacting with their environment. Another trend is increasing digitization (where developments are running parallel to those in automation, Eurofound, 2018), which means that machines in general and industrial machines in particular are increasingly being connected to each other and to internet via networks. With regard to these trends, it is vital not to lose sight of the future. Tomorrow's levels of machine safety must be defined today, to ensure that – from the design and development stage onwards – robots can be made inherently safe.

In this context, TNO (under a contract from – and in cooperation with – the Ministry of Social Affairs and Employment; hereinafter SZW) has, in recent years, investigated the extent to which European Directive 2006/42/EC, on machinery (hereinafter: the Machinery Directive), is future-proof. In previous reports, TNO has explored new risks and potential control measures. These range from linking industrial machines to internet, and to one another via internet (Steijn, Van der Vorm et al., 2016), the introduction of robots into the workplace (Steijn, Luiijf, et al., 2016), and humans and robots sharing the workplace (Jansen et al., 2018). The focus of this report will be the following research question, formulated by SZW:

With regard to machines equipped with machine learning, what essential health and safety requirements should be included in the Machinery Directive?

Developments related to industrial machines. Industrial machines are managed by control systems. In traditional industrial machines (TIMs), operations are scripted. This means that the TIM performs its prescribed (programmed) task within a structured environment (see Figure 1).

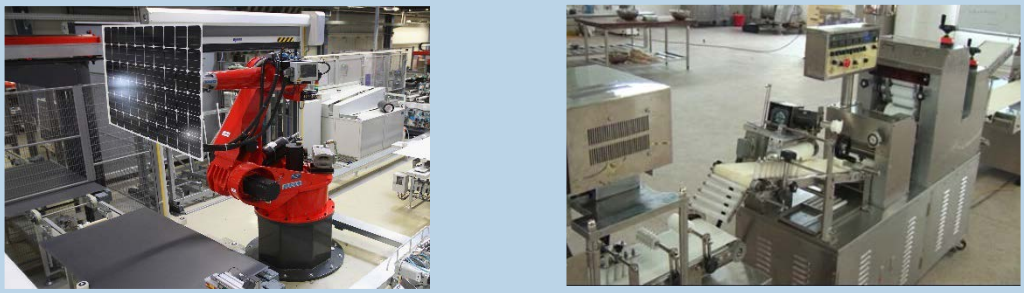

Figure 1[1] Some examples of traditional industrial machines (TIMs) with a fixed task. Often located in confined spaces and with no machine learning

[1] Figures from https://techcrunch.com/2017/03/19/y-combinator-has-a-new-ai-track-and-wants-startups-building-robot-factory-tech-to-apply/?guccounter=1 and http://dutch.foodmakingmachines.com/sale-4110180-industrial-automatic-dough-forming-machine-steamd-bun-making-machine.html

Developments related to industrial machines (continued)

Recently, more advanced industrial machines (AIMs) have been introduced into the workplace. These AIMs, which are also known as robots or cobots, are still working on defined tasks, in structured environments, and they can learn to perform new actions in this context (see Figure 2). Developments in machine learning have enabled AIMs to refine and improve their actions, based on new data. This eliminates the need to specifically program each and every aspect of the entire spectrum of required actions. Development work on AIMs focuses on making them less dependent on human operators. In line with this, the associated technical innovations are focusing on capabilities that make machines better able to recognize – and respond to – their environment (Eurofound, 2017).

Figure 2[2] Examples of advanced industrial machines (AIMs) capable of performing multiple tasks, that operate alongside people in the workplace, and that are equipped with machine learning technology.

In the future, further refinements to industrial machines (or FIMs, futuristic industrial machines) will include the real-time processing of information, problem solving, mobility, sensor systems, learning, and adaptability. In a recent study (Grace et al., 2018), AI experts stated there is a 50% chance we will see the development of AI that is more capable than humans on all fronts (High-level Machine Intelligence, or HLMI) within the next 45 years. These figures are based on estimates from 352 published scientists in the field of machine learning. In some limited areas, developments are expected even sooner. AIMs or FIMs capable of operating in unstructured environments (e.g. construction sites) are expected within 15 years (Robotics VO, 2013).

The Machinery Directive was implemented in the Netherlands via the Commodities Act on Machinery. The Commodities Act on Machinery refers to Annex I of the Machinery Directive. This includes essential health and safety requirements (H&S requirements). These requirements must be observed by manufacturers and designers, when placing their machinery on the market.

The Machinery Directive was drawn up at a time when there were not yet any practical applications for machines equipped with machine learning. The term ‘machine learning’ was coined by Arthur Samuel, in 1959. He researched the feasibility of enabling machines to learn from experience, thus eliminating the need to specifically program each and every aspect of their operations (Samuel, 1959). Machine learning is part and parcel of Artificial

[2] Figures from https://www.nbt.nhs.uk/about-us/building-brunel/automated-guided-vehicle-system and https://www.talentica.cz/robot-nebo-kobot/

Intelligence (AI), and of a machine's control system. The Machinery Directive contains H&S requirements that control systems are required to meet.
However, it is questionable whether these requirements will be sufficient to ensure safety when it comes to machines that are designed to become learning entities, capable of modifying their own behaviour.

The Machinery Directive's H&S requirements apply to manufacturers and developers of machines for both the professional context and for the consumer market. This report will mainly focus on industrial machines that use machine learning algorithms during their operational phase. This is in contrast to the use of machine learning algorithms to train machines during their design phase only. It is anticipated that machines capable of using machine learning to modify their actions in the workplace might pose new risks. Accordingly, the proposed H&S requirements listed at the end of this report will mainly apply to the former category of machines.

To answer the above research question, a desk study on this topic was carried out. Interview requests were sent to various experts in the field of robotics and machine learning. The information gathered in this way was used to identify several important topics relating to the preparation of supplementary H&S requirements. Based on this information, an initial proposal was submitted, concerning possible supplementary H&S requirements. These requirements were then tested, as part of the follow-up of two internal workshops held at TNO. A final version of these requirements was subsequently drawn up. Accordingly, this report consists of two parts. The first part concerns the initial survey. Part 2 gives further details of the validation. Below, there is a brief explanation of the Machinery Directive and of machine learning. This is followed by a description of the project's approach.

1.1 The Machinery Directive

Numerous product directives[3] have been published in the European Union (EU) since the 1980s. Such directives are also referred to as product safety directives, new approach directives or CE marking directives. In the Netherlands the Machinery Directive has been incorporated into Commodities legislation, more specifically into the Commodities Act on Machinery.

These directives have the following objectives:
1 Establishing an internal European market for products, by harmonizing the requirements imposed on such products.
2 Providing a high level of health and safety protection for those who work with/deal with such products and, where appropriate, for animals and the environment.
3 A level playing field for conformity assessment bodies (including notified bodies).

The health and safety requirements imposed on machines are important for machine manufacturers operating within the European internal market. This is included in Articles 2 and 5 of Directive 2006/42/EC on machinery (also known as the Machinery Directive). Any

[3] These product directives apply to the entire European internal market. This includes the EU Member States, EFTA Member States, and Switzerland (in the latter case, by means of a mutual recognition agreement or MRA).

manufacturers (or their authorized representatives) planning to place a machine on the market must comply with the Machinery Directive's requirements before the machine in question can be put into operation. How exactly they comply with these requirements is a matter for these manufacturers (or their authorized representatives). Based on the H&S requirements listed in Annex I of the Machinery Directive, manufacturers are required to:

1 List the hazards caused by the machine.
2 Eliminate these hazards.
3 If this is not possible, they must take measures to manage these hazards properly.
4 List the remaining hazards in the documentation they supply with the machine.

To this end, a risk assessment must be drawn up for the machine in question, to assess which Annex I health and safety requirements apply. The manufacturers then apply these requirements when constructing the machine in question.
Manufacturers can use the European harmonized standards to help them comply with these H&S requirements. The latter standards are formulated in more specific terms than the general health and safety requirements.
The manufacturer furnishes the machine with CE marking and a declaration of conformity. This indicates that the machine has been constructed in accordance with the essential health and safety requirements of the Machinery Directive.

Annex I, Section 1.2.1 of the Machinery Directive, contains H&S requirements for control systems. These H&S requirements will also apply to machines equipped with machine learning.

1. The control system can withstand the intended operating stresses and external influences.
2. Faults must not lead to hazardous situations.
3. Errors in the control system logic must not lead to hazardous situations.
4. Human errors during operation must not lead to hazardous situations.

1.2 Machine learning

Machine learning is a complex topic, and the field is still undergoing rapid development. We will not explore machine learning in any great depth here. However, we would like to briefly introduce the topic in connection with the objective, which is to formulate potential essential H&S requirements for machines equipped with machine learning.

Machine learning is part and parcel of the Artificial Intelligence (AI) research area. Within this research area, efforts are being made to replicate intelligent behaviour using computers. Machine learning enables machines to recognize patterns in complex data and to learn from experience (or, in this case, data). As a result, they are able to optimize their performance or their ability to execute certain tasks. In specific terms, this could involve optimizing the time required to carry out a task or minimizing the number of incorrect decisions taken, for example. Thanks to improved information processing, memory capacity and computing power (compared to humans), machines equipped with machine learning can be used for

tasks that are a) too difficult for people to carry out, b) too complex to program[4] and/or c) require flexibility[5] (Shalev-Shwartz & Ben-David, 2014).

There are various types and categories of machine learning. For the purposes of illustration, the topics we discuss here are Reinforcement learning (RL), Supervised machine learning (SML), and Unsupervised machine learning (UML)[6].

In reinforcement learning, the machine learns to make the right associations between input and output, based on positive and negative feedback. This enables the machine to optimize the required behaviour.

In the case of SML, the machine is offered examples (input) together with the corresponding required actions (output) (Russell, & Norvig, 2010). This enables the machine to learn what is 'right' and what is 'wrong'. The aim here is to generate a general rule about the output the machine must produce in response to a specific type of input. This is called 'supervised learning' because people are directly involved in the learning process, and because they can also test the process the machine has learned.

In the case of UML, a machine can 'independently' adapt to new situations. Based on input, the machine itself identifies patterns or structures, which it can then apply to new input. Each individual situation provides new input, and the machine uses this to discover patterns it can use. In this way, the machine itself links certain output to certain input, without human guidance.

It is important to note that, as things stand, the learning abilities of machines equipped with machine learning are still entirely dependent on their programming and algorithms. So, as yet, there is no such thing as a truly independent learning entity. What we have is a machine that, thanks to a more complex control system, is able to process a broader range of input and to optimize its output accordingly. Thus, for the time being, machines are always given objective functions (that the machine uses to calculate its objective).

Example: AlphaGo

Basically, AlphaGo (Silver et al., 2017) knew nothing about the game of Go. It learned the rules during a training phase, via SML, by receiving feedback as it played. Next, while in actual operation, the system improved every time it won or lost, by means of reinforcement learning. In this way, the latest version of AlphaGo has achieved superhuman performance, beating the human Go champion 100-0.

[4] This includes tasks that people perform 'naturally', such as speech recognition, and others that are too difficult for people, such as weather forecasts.

[5] Traditionally, machines have not been able to deviate from a fixed script. Now, machine learning enables them to respond to changing situations in their surroundings.

[6] See also: https://www.e-sites.nl/blog/476-machine-learning-een-korte-toelichting-op-de-techniek-en-toepassing.html

1.3 Organization of this report

The next chapter describes the method used to consult currently available literature, experts from the field, and scientific experts. This process generated insights into the new risks posed by machines equipped with machine learning. Chapter 3 briefly summarizes the results. The ultimate aim is to arrive at a draft proposal for supplementary essential H&S requirements for machines equipped with machine learning. This draft proposal is then tested internally, by a range of TNO experts from relevant fields. Details of the test method used are given in Chapter 4. Chapter 5 illustrates the main aspects to emerge from this test process. Based on these aspects, a final proposal has been established concerning essential H&S requirements for machines equipped with machine learning

2 Method - Part 1

The results in this part of the report were obtained by studying the available documentation. That included the Machinery Directive, interviews and a workshop involving experts from the field and scientific experts in the area of robotics. Each method used is briefly explained below. Chapter 3 illustrates the main aspects meriting particular attention that emerged during the desk study. They are integrated with the results of the interviews and of the workshop.

2.1 Desk study

The topic of this report is highly complex and extremely innovative. As a result, it was necessary to consult the literature (and the grey literature) to identify reference points for essential health and safety requirements. We also analysed the current version of the Machinery Directive.

2.2 Interviews

Semi-structured interviews were held with twelve experts in robot safety, human-machine interaction, cognitive engineering, artificial intelligence, ethics and legislation, as well as in the development, implementation and use of robots. The material covered in the interviews included the debate surrounding machines – in the form of robots – equipped with machine learning. That is why the word 'robots' will be used when discussing the results.

Table 1 Backgrounds of the interviewees

Job description	Specialization
Expert from the field	Technology, labour and privacy in the workplace
Expert from the field	R&D Robotics
Expert from the field	Machinery Directive expert
Scientific expert	Interfacing Law & Technology
Scientific expert	Intelligent Control and Robotics
Scientific expert	AI expert
Scientific expert	Specific focus on robots and AI
Scientific expert	AI expert
Scientific expert	Integrated Systems Engineering
Scientific expert	Intelligent man-machine systems
Scientific expert	Robot and AI ethics
Scientific expert	AI and Human-Computer Interaction (HCI) expert

2.2.1 Participant interviews

Based on the literature and on an internet scan, an actor analysis was carried out. This mainly involved actors with a knowledge of robot systems in general and of AI in particular. These experts were then contacted by phone or mail, and invited to an interview. The goal was to get at least ten of these individuals to take part.

A total of 32 invitations were issued in March. Table 1 includes a summary containing an anonymous description of those participants who ultimately attended interviews.

2.2.2 Interview protocol

The interviews were semi-structured, which means that a protocol was drawn up in advance. The questions contained in that protocol provided a guideline during the interview. During the interviews, the questions mainly focused deliberately on areas about which the interviewee in question had a great deal to say. The interviews each lasted from 60 to 90 minutes. Annex A contains details of the interview protocol used. Where necessary, the list of questions was tailored to the interviewee's background.

In addition to the protocol, the interviewees were given certain details in advance, to streamline the interview. This information was based on currently available literature and on the personal knowledge of TNO staff. To conclude, this information included a summary of three categories of industrial machines[7]:

Category 1: Traditional industrial machine (TIM)
These machines' programming enables them to automatically perform a simple task. These machines are often fixed in a specific location, and people are kept at a distance from them (e.g. by a safety cage). These machines cannot operate without direct human intervention (e.g. either by direct operation, or by supplying and removing the materials to be processed).

Category 2: Advanced Industrial Machine (AIM)
These machines are capable of performing multiple operations, or more complex operations. Some types can also exceed their original programming, thanks to machine learning. That means they can use data to carry out their tasks differently and more efficiently. These machines are no longer kept in a specific location, as such. Some of them can use sensors to 'see' – and respond to – their surroundings. They operate alongside people in the workplace. There is more 'cooperation' than 'operation', even if such cooperation is 'scripted'.

Category 3: Futuristic industrial machine (FIM)
These machines have an AI that approaches the level of human intelligence. They are creative, and can solve problems themselves. Rather than machines that are operated, they become 'agents' and, hence, robot colleagues.

2.3 External workshop

On 4 May 2018, the NEN's Industry & Safety platform held a 90-minute workshop entitled "*Cobots as an emerging risk in terms of occupational safety"*[8]. Thirty-two people (including two TNO project members) took part in the workshop.

[7] The interviewees were sent the following information (in tabular form) prior to the interview.
[8] https://www.nen.nl/Evenementen/Evenementdetailpagina/NENPlatform-Industrie-Veiligheid- 2.htm

The aim of the workshop was to answer the following questions:

1. What will it take to make future cooperation between users and machines (robots) in the workplace inherently safe?
2. To which set of human values (e.g. human privacy) should AI be attuned (in relation to taking decisions on the execution of the work, together with a human colleague) and what legal and ethical status should it have?
3. What modifications need to be made to the current legal framework (including the Machinery Directive) to effectively manage the risks associated with AI?
4. What is needed to facilitate constructive and healthy exchanges (including exchanges of knowledge) between all parties in the chain (manufacturers, system integrators, end users, AI researchers and policy makers), in the context of an effort to achieve inherently safe cooperation between humans and robots?

The workshop participants were divided into five groups of six people, who then brainstormed about the above questions. The result of each group's discussion was recorded on a flipchart sheet. In a plenary feedback session, one individual from each group presented the results of their discussion. The other participants were then given the opportunity to reflect on this. Questions 1 and 3 were directly related to this report's research question. The debate concerning these questions has been incorporated into the results (Part 1).

3 Results - Part 1

Based on information gathered using the methods described in the previous section, three topics emerged that are vital to the preparation and safeguarding of essential H&S requirements. The first of these was that a machine should not be at the top of the decision hierarchy. In other words, man must always have control over the machine, not the other way round. Secondly, great importance was attached to ensuring that the machines' algorithms and the data used were fully transparent. This applied not only to the operational process, but also to the development process. This transparency is required to guarantee quality. It is also needed to prevent the machines from developing a 'bias' or from becoming 'black boxes'. With the increasing complexity of the control systems involved and given that machines themselves can move unpredictably (based on what they learn), transparency is vital to prevent machines from becoming unpredictable 'black boxes'. Finally, with regard to safeguarding the H&S requirements, there must be clarity about how responsibilities are divided up between developers, integrators, and the end user, for example.

This chapter includes a further explanation of the above topics, based on the data collected. Based on this information, an initial version of the potential supplementary H&S requirements for machines equipped with machine learning is then proposed. The following chapters give details of the steps taken to turn a draft version into a final proposal for the H&S requirements.

The material covered in the interviews included the debate surrounding machines (in the form of robots) equipped with machine learning. That is why the word 'robots' will be used when discussing the results.

3.1 Human in command

Several of the interviewed experts seem to agree that robots should not be at the top of the decision hierarchy. A robot should above all be seen as a machine that can function in a decision support role, based on its properties. People must be able to choose whether (and how) they want to (or should) delegate decisions to machines, to achieve human-selected goals.

Furthermore, robots cannot be used in a way that requires them to independently make decisions affecting people in their environment. This is the case, for example, with 'trolley problems'. Trolley problems involve morally difficult considerations in dilemmas about people's lives, in which every choice has a fatal outcome (for example, the choice between saving someone you know personally, and saving five strangers). It is not inconceivable that autonomous machines will face such dilemmas in practice. This is already the case for automated guided vehicles, for example[9]. Ideally, it will always be possible to trace the responsibility for such choices back to people (e.g. the way in which an algorithm is programmed).

[9] Algemeen Dagblad newspaper (2018), Essay, man and his machine, we have long accepted the idea of robots killing people, 25 March 2018.

To this end, in its Artificial Intelligence private member's bill (EESC, 2017), the European Economic and Social Committee has called for a **human-in-command** approach to AI. This would be subject to certain preconditions for the responsible, safe and useful development of AI. Here, machines would remain machines, and people should and would always retain meaningful control over these machines. A basic condition for meaningful human control is that machines must be transparent to those users in the workplace who work with them. Further, there must be transparency to facilitate supervision of the process and any subsequent checks (e.g. during an incident analysis). In the next section, we will explore these forms of transparency in greater detail.

3.2 Ensuring transparency regarding algorithms and machine behaviour

Machines equipped with machine learning introduce new risks into the workplace. This is due to decreased transparency concerning what they do and why they do it. The robot's control system is in danger of becoming a 'black box', as previously described. This makes robots' operations less predictable for those users who work with them. It also makes it more difficult to retrospectively determine why a robot performed a given action. In terms of legislation and regulations, the following three types of transparency must be guaranteed if we are to have truly safe robots:

- Transparency of the data sets and algorithms used (to prevent bias);
- Transparency of the machine's algorithms (to avoid a 'black box' scenario);
- Transparency of the development process (to guarantee quality).

Pitfalls in the pursuit of transparency
Weller (2017) identified various pitfalls in the pursuit of transparency, which need to be prevented. Firstly, transparency can be misleading if it also results in some information being withheld. Secondly, increased transparency can cause some individuals' privacy to be compromised. Thirdly, it is important to avoid situations in which transparency is seen as an end in itself, and where this acts as a brake on innovation. Efforts to increase transparency must always be motivated by the need to improve safety. Finally, greater transparency can lead to discriminatory behaviour (by providing reliable information about certain personal characteristics, such as ethnicity).

3.2.1 Transparency of the data sets and algorithms used to prevent bias

Robots must always be 'explainable'. In other words, it must always be possible to trace their behaviour and decisions back to programmed algorithms. Accordingly, it is important for the algorithms and the data that enable robots to function to be transparent and understandable. In this way, you can avoid situations in which a robot's operations (correct or incorrect) can no longer be traced to specific causative factors.

One of the experts cited the example of a research project at Microsoft, which involved the development of an autonomous system based on machine learning algorithms. This system could be used to support doctors, when assigning patients to a given risk category (high or low). Chest pains and symptoms of asthma were designated as low risk, but this classification turned out to be the result of bias.
The reason was that this particular group of patients often tended to be admitted at an early stage, which resulted in a low risk of mortality.

If people are to intervene (to prevent 'potential bias' or 'trade-offs'), they must have an understanding of the algorithm involved, of which datasets have been used, of the underlying hypothesis, and of how the various performance metrics (on which a diagnosis is based) operate.

Translated into the everyday work situation, inappropriate testing of the algorithm that controls the robot could lead to a situation in which a welding robot starts welding before users in the area have put on their protective goggles. The robot may have mistakenly learned that welding was a low-risk operation, as there had never been any incidents (because, previously, the users had always been wearing protective goggles while the robot was welding).

Another expert cites the example of a Chinese study, in which an autonomous system decided who had committed a crime and who was innocent. Some of the photographs used in this study were taken from people's LinkedIn profiles, while other photographs were of actual detainees. This resulted in a bias, because people use LinkedIn to further their career, so of course they use a photograph that shows them smiling. Those who are sentenced to prison, on the other hand, tend to look angry rather than happy. Thus, if this characteristic is one of the selection features on which a decision is based, then the decision framework is incorrect and it cannot be reliably used to identify offenders.

Another scientific expert stated that learning machines designed to make real-time decisions depend on factors such as time, their perception, and their interpretation of their surroundings. In the case of autonomous drones that are required to record some aspect of their environment, for example, no two landscapes are completely alike. In short, certain factors in the environment can cause bias, which may influence the reasoning that precedes a decision. The environment in which a system has been tested is often not the same as a real world environment. Accordingly, systems that have passed laboratory tests but which have not been subjected to external exploratory testing pose risks when placed in an industrial setting. This is because a test environment is usually unable to take full account of the context in which the final practical application will operate.

Bias can also occur while a robot is being programmed. For instance, programmers may – subconsciously – incorporate their personal perceptions and experiences into the algorithm. It is also important to prevent robots developing a bias. This can be done by drawing on good data sources, and by ensuring that data sources are free from external influences. To this end, for example, the European Economic and Social Committee has called for the establishment of a European AI infrastructure (in the form of an Artificial Intelligence private member's bill; EESC, 2017). The plan involves open source and privacy-respecting learning environments, lifelike test environments, and high-quality datasets for developing and training AI systems.

Bias can also result from external influences (e.g. cyber criminals or hackers) or software errors. The interviewees had various suggestions on how to deal with this. One idea involved the use of supporting software to assess whether an algorithm is being implemented correctly ('correct by design'). Another option is to program an extra layer into the software, one that checks and regulates the software's own behaviour. This could involve a decision-making framework (as provided for in the statutory H&S requirements contained in the

Machinery Directive), for example. A '*safety loop*' in the software will then deactivate the robot if its planned behaviour does not meet the safety criteria.

3.2.2 Transparency of the machine's algorithms during operation – to avoid a 'black box' scenario

As stated above, with non-transparent datasets and algorithms there is a threat that robots will become 'black boxes' (in the negative sense of the term). These are closed systems. They receive input and produce output, without explaining how and why they decided on that particular output. There are several potential drawbacks here:

- Any bias that arises in the robot may go unnoticed (see also Section 3.2.1).
- It is impossible to deduce whether the robot's decisions are based on fair criteria (any results that are incorrect or that were derived by non-scientific means are invalid).
- The sources of system errors are more difficult to trace, which means they are more difficult to correct.
- The robot can be hacked either internally or from outside the system, which can pose a risk to the integrity of its software.

This year, the European Commission will address the topic of 'algorithmic transparency'. It has proposed a three-pronged approach, in relation to AI (2018). This does not involve disclosing an algorithm's source code, as such. It can take various forms, depending on the situation. This includes meaningful explanations (as required by the General Data Protection Regulation (GDPR)[10] with regard to automated decisions based on personal data, for example). Another involves reporting to the competent authorities (as required in the Markets in Financial Instruments Directive (MiFID II)[11]).

In addition, the European Economic and Social Committee has advocated the development of a standardization system (through an 'Artificial intelligence' private member's bill; EESC, 2017) for the verification, validation and control of AI systems. This system should be based on a broad spectrum of standards in the areas of safety, transparency, comprehensibility, explicability and ethical values.

The Future of Life Institute also became involved, when it launched its 21 Asilomar AI Principles at the 2017 Asilomar conference. With regard to transparency, they emphasize that this must extend to failures. If an AI system causes damage, then it must be possible to trace the cause. One of the interviewed experts indicated that, with this in mind, it would make sense for the robot to have a 'black box' in the positive sense of the term. This would operate like an aircraft's black box, which is used to analyse any failures, after the event.

In order to manage any risks to users, robots must also make their intentions clear, through their interface. For instance, a robot intending to move could make this clear by using a LED light to indicate the direction in which it is going to move. Another example of how to interpret intentions involves plotting the predicted positions on the operator control interface.

[10] For more information about the GDPR see: https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en

[11]: For more information about the MiFID see: https://ec.europa.eu/info/business-economy-euro/banking-and-finance/financial-markets/securities-markets/investment-services-and-regulated-markets-markets-financial-instruments-directive-mifid_en.

This shows the operators where the robot will be in the future, enabling them to anticipate any associated effects.

Weller (2017) provides several examples of information feedback that robots could generate to improve transparency, for example. This could take place at different levels. Developers want to be able to see whether the system is operating well or badly, and to identify the causative factors involved. The same thing applies to users. Users want to know what the robot is going to do and why. This helps them to become more familiar with the robot, in the event of future (unforeseen) situations. Users or maintenance engineers also want to know how the system arrives at a given prediction or decision. They also want to be able to check whether the underlying system is still operating according to the legal requirements of the Machinery Directive and the design specifications of the manufacturer and customer. Data could also be stored, enabling experts – such as the algorithm's developers – to make improvements. Alternatively, in the context of an incident analysis, scientific experts could use this data to deduce why something went wrong.

3.2.3 Transparency of the development process to guarantee quality

In addition to the importance of transparency in a robot's data and algorithms during operation, it is important for transparency to be generated as part of the process (i.e., from design to prototype) by which a robot's control algorithms are developed. One of the scientific experts explained that, while developers often use a range of tests to assess a robot's performance (this could involve both the performance of a task and compliance with the requirements of the Machinery Directive), they do not always publish details of tests in which the robot performed poorly. This allows developers to present a better picture of their robot's performance.

One well-known way of manipulating research results is probability-hacking (P-hacking). P-hacking is a term that owes it origins to a battle raging in the area of statistics (Head, Holman, Lanfear, Kahn, & Jennions, 2015). It is related to the standard used to express significance – the p-value. Evidence from the social sciences suggests that, when studies are repeated, the results obtained are often very different from those obtained the first time round. While P-hacking is primarily a topic of interest within the social science domain, it is also important in other domains (e.g. when testing technical specifications for safe designs). Being keen to innovate and stay ahead of the competition, researchers could go on repeating a study, using different standards each time, until they get a result that is good enough to be published. As a check on such practices, all studies must satisfy a replicability requirement. In other words, other researchers must be able to exactly replicate a study, to check its results.

If robots defined as safe and reliable – based on P-hacking – were to enter everyday use, this would pose a threat to people's safety. However, it would be very difficult to determine whether any specific instances of damage, arising from unsafe situations, were related to P-hacking. This is because developers are unlikely to publish details of any previous tests in which their robots performed poorly. Only the results of tests 'demonstrating' that the robot performs well will be published. This gives rise to a distorted picture of the robot's quality. In everyday practice, the goal of success has a profound influence on P-hacking. Developers naturally want to present their robot in the best possible light. At the same time, there is very little risk of being caught, as P-hacking is difficult to detect.

The medical sector avoids such situations by imposing a pre-registration requirement on new drugs, concerning research into their efficacy. This shows whether publications in which an effect was found (perhaps by chance) were preceded by a number of studies in which no such effect was detected. As a result, this reduces the risk of ineffective (or even harmful) drugs entering the market.

In the area of robot development, there is currently no legal requirement for the pre-registration of research designs. Hofman, Sharm and Watts (2017) impose various requirements on the design of studies in which the performance of robots is determined. The goal of these requirements is to ensure that such studies remain transparent, making it easier to identify any research results that were found by chance. In addition, these requirements enhance the studies' replicability.

The pre-registration of research designs would create openness with regard to the processes involved in developing a robot's machine learning capabilities. This approach makes the details of an algorithm's history (including its development history) and performance clear to other researchers. The points that need to be clarified during pre-registration of the research design include:

- The type of datasets used.
- The relationship between training data and validation data.
- How frequently testing took place, and what kind of sample data was used for this purpose.
- How the hypothesis (about how the machine can learn most effectively) was developed.
- All pre-processing choices – in addition to the choice of data, this concerns the way data is cleaned up, how it is labelled, and the range of potential labels (or alternative labels).
- The types of algorithms used, or whose use is planned.

The ultimate goal of these requirements is to ensure that applications actually fulfil the intentions and claims of those who place them on the market.

3.3 Responsibilities

Many parties are involved in the process that commences with the development of a robot and ends with its ultimate application in practice. These include the developers of different robot parts, the integrator who installs the robot parts at the customer's site, and the customer who uses the robot.

In practice, it is not always clear where certain responsibilities lie, with regard to ensuring that the final product meets – and continues to meet – all necessary requirements. Until such time as the confusion surrounding this issue is resolved, any supplementary legal requirements will not produce the desired effect. The risks (or potential risks) inherent to AI systems are currently the focus of extensive planning and risk mitigation work. This work must be proportionate to its anticipated impact (proportionality).

Information gathered during the interviews shows that responsibility must be covered at chain level. The manufacturers must ensure that their product or service complies with H&S requirements.

3.4 Draft proposal concerning supplementary essential H&S requirements for machine learning

Based on the above information, the following potential supplementary H&S requirements were formulated (in addition to the existing H&S requirements for control systems). These requirements are subdivided per topic, as discussed above, with a reference to the relevant section from Annex I of the Machinery Directive (MD).

3.4.1 Human in command

When developing machines equipped with machine learning, the key principle should be to ensure that people are always in control of the situation (and of the machine). In this context, the important key principles are as follows:

- People give orders to the machine, the machine does not itself give orders to people.
- People must always be able to intervene and 'overrule' the machine safely.
- The man-machine interface must be designed to ensure that people are always aware of what the machine is going to do.

In the case of machines equipped with machine learning, it is also important to define the physical environment in which the robot can be used safely (see text box entitled *Framing a machine's learning potential*). Also, with regard to the hardware, it is of paramount importance that the machine's surroundings be structured in a way that enhances the robot's functionality, while safeguarding the health and safety of users.

Supplementary H&S requirements:

- Machines equipped with machine learning technology may not be placed or installed in situations where they themselves are required to make assessments concerning injuries to people and/or damage to their surroundings (supplementary to: Safety and reliability of control systems; Section 1.2.1, Annex I MD).
- Machines equipped with machine learning technology must be equipped with an emergency stop function, so that they can be deactivated/overridden at any time. Once the machine has been deactivated, the situation is safe (supplementary to: Stopping, Section 1.2.4, Annex I MD).
- Machine learning must not cause the machine to exhibit new behaviour that exceeds its defined task and working environment (supplementary to: Safety and reliability of control systems; Section 1.2.1, Annex I MD).

Framing a machine's learning potential

One way to eliminate the risks of a machine equipped with machine learning becoming unpredictable is to frame its learning potential. In traditional industrial machines, the machine's working range is defined (see Figure 3, in Section IV: Chapter 4, OSHA Technical manual (OSHA, undated)).

The machine's working range is relatively limited (black) because it repeats the same movement. Its maximum potential range (white) is much larger, however. In the case of machines equipped with machine learning, their entire maximum potential range is their working range. This means that there is less safe space around the machine for users (in Figure 3 the safe area is everything outside the striped section), because the machine's specific actions are no longer fixed.

This risk is managed by clearly defining and, if necessary, restricting the working range (black) within which the robot can perform new actions.

The area that is restricted for users will then have to be adjusted accordingly. In this situation, the robot will still be able to modify its behaviour, however its unpredictability will be limited, as its behaviour will be framed and defined.

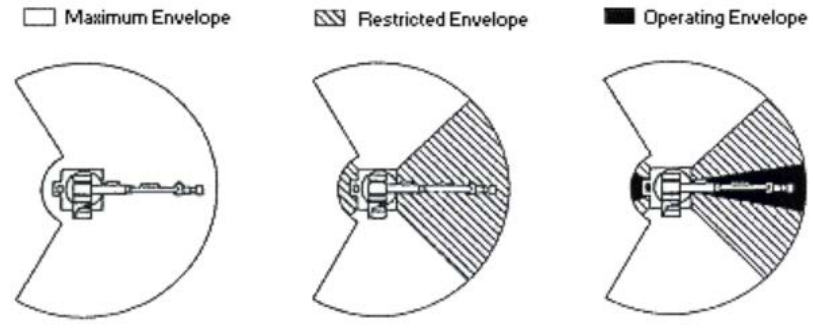

Figure 3 Visualization of the defined movement space of an industrial machine. From left to right – maximum movement space (white), inaccessible movement space (striped) and used movement space (black)

3.4.2 Transparency

Due to the increasing complexity of control systems, and to developments in the field of machine learning (and AI), industrial machines' control systems are at risk of becoming black boxes (Figure 4). This may also have been true of traditional industrial machines. However, in that case you could be sure that the same input would produce the same output, unless something was wrong with the machine. This can be problematic, however, if the machine in question is capable of deviating from its original programming, as the same input could then produce a different output. This could happen if the machine were to assess input in a different way. Accordingly, steps must be taken to ensure that, in machines with learning capacity, the process between input and output is overtly transparent.

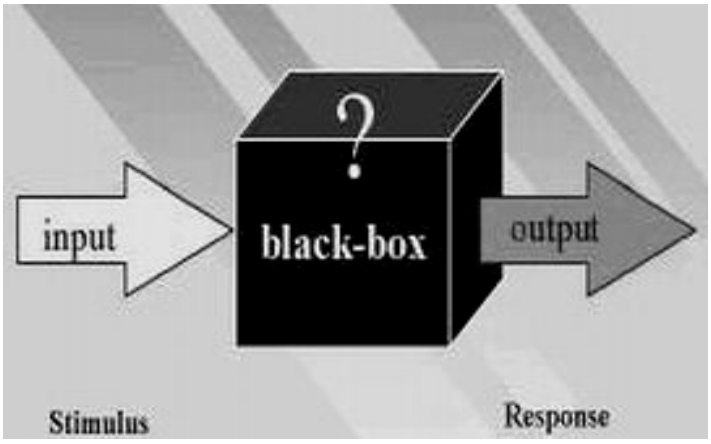

Figure 4[12] It is unclear how the output is determined on the basis of the input.

On the one hand, the machine must become more transparent to users in the workplace. This can be achieved by communicating details of the current movement pattern and of any potential departures from it. This could, for instance, involve routine feedback (at the end of the day) concerning everything the machine has learned, and whether this is in compliance with the H&S requirements in force at the time. On the other hand, the software must be readable at all times: This could be used to deduce the basis for the current movement pattern and how a given incident could have occurred. To this end, it is important that the development process is already transparent concerning:

- The original code for the algorithm.
- The raw data used for the algorithm.
- The actions the machine performs, based on the data and the algorithm.

The purpose of a type approval is to show which requirements have been imposed and which have been met, before datasets and algorithms can be described as transparent. In practice, this means that designers, builders, and manufacturers who produce machines equipped with machine learning should submit details of their machine or production process to a conformity assessment body (See Regulation 765/2008 Article 2 (13)) before placing this product on the market. These bodies calibrate, test, certify and inspect the machine's components.

3.4.2.1 Transparency of the machine's algorithms during operation

Supplementary H&S requirements:

- In view of its role, a machine equipped with machine learning technology must be able to respond to people adequately and appropriately (verbally through words and/or non-verbally through gestures, facial expressions or body movement; supplementary to Ergonomics; Section 1.1.6, Annex I MD).
- A machine equipped with machine learning technology must be able to communicate its intentions (what it is going to do and why) to users in a comprehensible manner (supplementary to: Ergonomics; Section 1.1.6, Annex I MD).

[12] Image taken from https://tvtropes.org/pmwiki/pmwiki.php/Main/BlackBox

- Machines equipped with machine learning technology should assist users by distinguishing the part of their analysis that is based on 'supervised learning' from the part that is based on 'unsupervised learning', in order to make the foundations of the analysis transparent (supplementary to: Safety and reliability of control systems, Section 1.2.1, Annex I MD).

3.4.2.2 Transparency of the datasets and algorithms used in the control system

Supplementary H&S requirements:

- The actions of a machine equipped with machine learning must be traceable (in advance and retrospectively), based on transparency in the datasets used, as well as the test environments (incl. scenarios used in the algorithm's training and validation models) and the decision frameworks or assessment criteria for algorithm-based decisions (supplementary to: Safety and reliability of control systems, Section 1.2.1, Annex I MD).
- Any decisions made by a machine equipped with machine learning technology must be logged and retained (supplementary to: Safety and reliability of control systems, Section 1.2.1, Annex I MD).

3.4.2.3 Transparency during the development process

Supplementary H&S requirement:

- The use of machine learning technology must be restricted to systems that have undergone public pre-registration of the study's research design – including the corresponding machine performance report (supplementary to: Safety and reliability of control systems, Section 1.2.1, Annex I MD).

3.4.3 Responsibilities

No supplementary H&S requirement is proposed with regard to this topic.

4 Method - Part 2

Two internal working sessions were held at TNO (on 31 July and 2 August 2018) to validate the proposed H&S requirements, described above. In both working sessions, TNO experts were questioned about various aspects (comprehensiveness, applicability, necessity) of the H&S requirements that had been identified.

Table 2 Individuals participating in the working sessions and their specialization

Job description	Specialization
TNO expert 1	Neural networks, deep learning and intelligent imaging.
TNO expert 2	Strengthening Visual qualities in Robotic systems.
TNO expert 3	Cybersecurity & IT
TNO expert 4	Psychometrics and statistics
TNO expert 5	Artificial General Intelligence & Ethics
TNO expert 6	Behavioural modelling and algorithm testing

The participants were given various items, such as the first version of the proposed H&S requirements, sorted by the relevant section of the Machinery Directive (see text box on next page). During the working session, the proposed H&S requirements' completeness and importance were assessed. At the end of the working session, the participants were asked to score the H&S requirements on the basis of their relevance:

0 = not important;
1 = somewhat important;
2 = important;
3 = very important.

Essential H&S requirements sent to working session participants

Current essential H&S requirements for control systems (goal-oriented requirements):

- H&S requirement 1: The machine's control system can withstand the intended operating stresses and external influences.
- H&S requirement 2: Faults in the machine's control system must not lead to hazardous situations.
- H&S requirement 3: Errors in the control system logic must not lead to hazardous situations.
- H&S requirement 4: Human errors during operation must not lead to hazardous situations.

Supplementary essential H&S requirements (goal-oriented requirements):
With regard to control system ergonomics (Section 1.1.6, Annex I MD).

- H&S requirement 1: In view of its role, a machine equipped with machine learning technology must be able to respond to people adequately and appropriately (verbally through words and/or non-verbally through gestures, facial expressions or body movement).
- H&S requirement 2: A machine equipped with machine learning technology must be able to communicate to users – in a comprehensible way – what it is going to do and why.

With regard to the safety and reliability of control systems, (Section 1.2.1, Annex I MD).

- H&S requirement 3: Machines equipped with machine learning technology may not be placed or installed in situations where they themselves are required to make assessments concerning injuries to people and/or damage to their surroundings.
- H&S requirement 4: Machine learning must not cause the machine to exhibit new behaviour that exceeds its defined task and working environment.
- H&S requirement 5: Machines equipped with machine learning technology should assist users by distinguishing the part of their analysis that is based on 'supervised learning' from the part that is based on 'unsupervised learning', in order to make the foundations of the analysis transparent.
- H&S requirement 6: The actions of machines equipped with machine learning technology must be traceable (in advance and retrospectively), based on transparency in the datasets used, of the test environments (incl. scenarios used in the algorithm's training and validation models) and of the decision frameworks or assessment criteria for algorithm-based decisions.
- H&S requirement 7: Any decisions made by a machine equipped with machine learning technology must be logged and retained.
- H&S requirement 8: The use of machine learning technology must be restricted to systems that have undergone public pre-registration of the study's research design – including the corresponding machine performance report.

With regard to stopping (Section 1.2.4, Annex I MD).

- H&S requirement 9: Machines equipped with machine learning technology must be equipped with an emergency stop function, so that they can be deactivated/overridden at any time. Once the machine has been deactivated, the situation is safe.

5 Results - Part 2

After the working sessions, the participants were asked to score the proposed H&S requirements, based on their importance in terms of the health and safety of the users working with these machines. All of the scores awarded are shown in the table below. The level of importance is indicated by the sum of all scores awarded, in the last column.

These figures clearly show that two H&S requirements are considered to be less important than the rest. These are H&S requirement *5 Distinguish between SML and UML* and H&S requirement *8 Pre-registration of research*. In the next section, we will explore the information underpinning the assessments of the defined H&S requirements in greater detail. We will also examine its ramifications, with regard to the final proposals for H&S requirements.

Table 3 assessment of the proposed essential H&S requirements

Essential H&S requirements for control systems (Annex I, Section 1.2.1 MD)	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6	Total
H&S requirement 1: Stresses and influences	2	2	3	3	3	3	16
H&S requirement 2: Fault	3	3	3	3	3	3	18
H&S requirement 3: Error in logic	3	3	2	3	3	3	17
H&S requirement 4: Human error	3	3	3	3	3	3	18
Proposed supplementary essential H&S requirements for machine learning							
H&S requirement 1: Responding to people	2	1	3	3	2	3	14
H&S requirement 2: Communicating intentions	3	3	0	3	2	3	14
H&S requirement 3: No assessments concerning injury/damage	2	2	3	3	3	1	14
H&S requirement 4: Limitation of new behaviour	2	2	3	3	3	1	14
H&S requirement 5: Distinguish between SML and UML	0	0	0	0	1	2	3
H&S requirement 6: Transparency and tracing behaviour	2	1	0	3	3	1	10
H&S requirement 7: Transparency of decisions	2	3	3	3	2	2	15
H&S requirement 8: Pre-registration of research	1	1	*	*	3	1	6
H&S requirement 9: Emergency stop function	3	3	3	3	3	1	16

0 = not important;
1 = somewhat important;
2 = important;
3 = very important;
*No score awarded.

5.1 Final proposal concerning supplementary essential H&S requirements for machine learning

5.1.1 H&S requirement 1: Responding to people

The H&S requirement – as proposed in the draft version – contained a list of random examples that is not exhaustive. Accordingly, it is not necessary to cite these for a goal-oriented requirement. For this reason, these examples have been removed.

Definitive version of supplementary H&S requirement:
Machines equipped with machine learning technology must be able to respond to people adequately and appropriately.

5.1.2 H&S requirement 2: Communicating intentions

According to the participants, communicating the machine's intention (what it is going to do) is a minimum requirement and is already technically feasible. According to some of those who participated in the working sessions, well-programmed systems will also provide an explanation of why they are performing a particular operation, as this is a logical extension of what the system is going to do. Indeed, others who participated in the working sessions argued that once machines start explaining 'why' they perform a particular operation, it will no longer be possible to use deep learning (in view of the method's complexity) and – by extension – certain machine learning applications. This group of participants feels that machines capable of communicating why they are going to perform a certain action are technically beyond our grasp.

This report has, to a great extent, focused on why transparency is an important element for machines equipped with machine learning, in terms of guaranteeing user safety. Accordingly, it was decided that the 'why' question should be retained in the H&S requirement. This may mean that some forms of machine learning cannot be used in industrial machines until they are better understood.

Furthermore, the discussion revealed that use of the word 'why' can lead to confusion. For example, one of the points discussed was that if machines equipped with machine learning are to really understand *why* they are performing a particular operation, then advanced AI is required. Nor indeed, was the H&S requirement intended to be interpreted in this way. Accordingly, the requirement has been modified to exclude the word 'why'. Instead, it specifically states that machines must provide details of the information on which their actions are based.

Definitive version of supplementary H&S requirement:
Machines equipped with machine learning technology must indicate which actions they are about to perform and must provide details of the information on which these actions are based.

5.1.3 H&S requirement 3: No assessments concerning injuries to people

If man and machine can come into contact with one another, such assessments will inevitably be made. The effect of this requirement, in its original form, is that machines must be kept separate from people. However, the real purpose of the requirement is to ensure that machines do not make the final decision on ethical issues. Accordingly, the H&S requirement

has been reformulated to make it clear that such ethical decisions must not be left to machines.

However, it is acknowledged that situations in which decisions have to be taken 'on the fly' still pose something of a dilemma in this regard. According to some who participated in the sessions, the key principle must be that – in situations of this kind – the ultimate outcome is based on whatever society regards as tolerable, and not something that has been determined by an algorithm.

Definitive version of supplementary H&S requirement:
Machines equipped with machine learning are not permitted to make decisions or assessments in relation to injury to people or damage to the surroundings.

5.1.4 H&S requirement 4: Limitation of new behaviour

In addition to people, the draft H&S requirement includes the 'surroundings'. This requires further discussion about what exactly constitutes a defined working environment. Furthermore, according to those who participated in the working sessions, it would be better to replace the word 'behaviour' with 'actions'.

Definitive version of supplementary H&S requirement:
Machine learning must not cause the machine to exhibit new actions that exceed its defined task and working environment.

5.1.5 H&S requirement 5: Distinguish between SML and UML

According to those who participated in the working sessions, it is indeed possible to distinguish between decisions based on SML and those based on UML. However, it is not particularly useful to do so. Even if you have a good definition, there is nothing to be gained by knowing whether it is SML or UML. It is not relevant to the end user. In essence, to make corrective measures possible, it is important to identify the part of the algorithm where something went wrong. This is covered by requirement 7. Thus, it is certainly a good idea to incorporate corrective measures for the decisions.

Definitive version of supplementary H&S requirement:
If they take incorrect decisions, machines equipped with machine learning technology must be retrospectively correctable, to prevent any future recurrences of that particular error.

5.1.6 H&S requirement 6: Transparency and tracing behaviour

To trace the machine's behaviour, methodological transparency is needed. A number of those who participated in the working sessions see this as an extreme requirement, given that it is even difficult for laboratories to define this in terms of ISO standards (hence the low scores awarded by some experts, Table 3). It was also noted that if a robot learns something independently (or via reinforcement learning), you cannot check and/or correct the outcome (if the data is not shared). The research team decided that, while it will be difficult to comply with the H&S requirement, this in no way detracts from this requirement's importance in terms of making machines equipped with machine learning inherently safe. Thus the research team decided to include it anyway. The word 'behaviour' is a gain replaced with 'actions'.

Definitive version of supplementary H&S requirement:
The actions of a machine equipped with machine learning technology must be traceable in advance and retrospectively, based on transparency in the datasets used, as well as of the test environments[13] and of the decision frameworks or assessment criteria for algorithm-based decisions.

5.1.7 H&S requirement 7: Transparency of decisions

No comments. The participants were agreed.

The fact that the machine and its parts can be audited is an important condition. If the datasets used by the machine are recorded in a system, then it will also be possible to audit the machine itself. The decision-making processes of machines equipped with machine learning must also be logged, and retained for auditing.

Definitive version of supplementary H&S requirement:
The decision-making process of a machine equipped with machine learning technology must be logged and retained[14].

5.1.8 H&S requirement 8: Pre-registration of research

Those who participated in the workshop approve of this requirement's underlying intention – to use scientific evidence to demonstrate that the machine is safe. However, a requirement like this poses considerable difficulties for small and medium-sized enterprises (SMEs). For the time being, organizations are not obliged to release their datasets, so their intellectual property will remain protected. Various alternative methodological verifications could also be used to guarantee quality. However, these are not discussed in detail in this report.

The participants could not agree whether the pre-registration of datasets should be included as an H&S requirement. All the more so because, according to the participants, this is mainly a means-oriented requirement to guarantee the transparency of the datasets and algorithms used. For this reason, our research team decided that, while the pre-registration of research is certainly an important means of guaranteeing the quality of new designs being placed on the market, it is not necessarily an H&S requirement. In other words, the necessity or duty of introducing pre-registration will have to be conveyed in a different way. Accordingly, this H&S requirement was not included in the definitive set of H&S requirements.

5.1.9 H&S requirement 9: Emergency stop function

According to those who participated in the working sessions, this requirement is feasible because the machine is simply switched off using a 'hard wired' safety circuit. According to a number of participants, however, this adds nothing new to machine learning as such. This is because it also applies to devices that do not involve machine learning (e.g. sensors), as described in Annex I, Section 1.2.4.3.) of the Machinery Directive. This requirement can, therefore, also be dispensed with in the new (supplementary) essential H&S requirements.

[13] For example, scenarios used in training and validation models for the algorithm
[14] In such a way that this information remains available for a minimum period (which is yet to be determined). It could then be checked during audits or incident analyses, for example.

5.1.10 Comments relating to current essential H&S requirements for control systems

The first, current H&S requirement for control systems states that the control system must be able to withstand external influences. In the context of machines equipped with machine learning, however, external influences are actually required. Here, it would be useful to specify that, in the case of machines equipped with machine learning, this also involves undesirable external influences.

Proposed amendment to the H&S requirement: The machine's control system can withstand the intended operating stresses and undesirable external influences.
Additionally, it can generally be stated that, if any errors or unforeseen conditions should occur in the control system, the machine should ideally revert to a safe state. As an additional barrier, there could be a requirement that only someone authorized by the organization to do so (e.g. the maintenance engineer) would be able to release the machine again.

5.2 Summary of final proposal

Current H&S requirements for control systems (Annex I, Section 1.2.1 MD):

1. The machine's control system can withstand the intended operating stresses and *undesirable* external influences.
2. Faults in the machine's control system must not lead to hazardous situations.
3. Errors in the control system logic must not lead to hazardous situations.
4. Human errors during operation must not lead to hazardous situations.

Supplementary H&S requirements:

With regard to control system ergonomics (supplementary to: Section 1.1.6, Annex I MD).

1. Machines equipped with machine learning technology must be able to respond to people adequately and appropriately.
2. Machines equipped with machine learning technology must indicate which actions they are about to perform and must provide details of the information on which these actions are based.

With regard to the safety and reliability of the control systems (supplementary to: Section 1.2.1, Annex I MD).

3. Machines equipped with machine learning are not permitted to make decisions or assessments in relation to injury to people or damage to the surroundings.
4. Machine learning must not cause the machine to perform actions that exceed its defined task and movement space.
5. If they take incorrect decisions, machines equipped with machine learning technology must be retrospectively correctable, to prevent any future recurrences of that particular error.
6. The actions of a machine equipped with machine learning technology must be traceable in advance and retrospectively, based on transparency of the datasets used, as well as of the test environments and of the decision frameworks or assessment criteria for algorithm-based decisions.

7. The decision-making process of a machine equipped with machine learning technology must be logged and retained.

5.3 Considerations

The following considerations have emerged during this project. However, they are not directly relevant with regard to answering this report's research question. Nevertheless, these considerations do provide an important context, within which the above H&S requirements must be considered.

5.3.1 Consideration 1

If the above-mentioned new H&S requirements for machines equipped with machine learning were to be included in the Machinery Directive, it would also be necessary to determine which conformity assessment procedure (or procedures) is (or are) applicable. This part (selection of conformity assessment procedure) is not covered by the present research.

Details of the various conformity assessment procedures (eight modules) can be found in Decision 768/2008/EC[15]. In essence, this involves three types of procedure.
1. The key principle is that manufacturers perform the conformity assessment of the design as well as of the machine's production themselves, as far as possible (Module A - Internal production control).
2. Another option is a conformity assessment procedure based on a type approval (module B, supplemented modules C, D, E and F.). Here, both the machine's design and the finished product are assessed for conformity with the H&S requirement. An accredited conformity assessment body must be involved in such conformity assessment procedures.
3. Finally, there is the option of a quality-assurance based conformity assessment (Modules D, E and H.) In addition, manufacturers must implement quality assurance systems that are compatible with relevant international quality standards (ISO 9000 and ISO 9001). Here too, the involvement of an accredited conformity assessment body is a mandatory requirement.

5.3.2 Consideration 2

In this report, we have used the example of pre-registration as a means of guaranteeing the quality of machine-learning software. As an alternative to pre-registration, the option of obtaining type approval for the machine learning software in question may be worth considering. As indicated under Consideration 1 (Section 5.3.1), the purpose of a type approval is to assess the design of the machine for conformity with the H&S requirements. Next, the finished product must also be assessed for conformity with the H&S requirements. Type approval can also be obtained for the software used in machines equipped with machine learning. Type approval would then be an option for guaranteeing the transparency of the software (see: Decision 768/2008/EC). The software's type approval can then be used in the later stages of the production process to demonstrate that the installed software is compliant with the software that was awarded a type approval. Alternatively, at the unit

[15] Decision No. 768/2008/EC of the European Parliament and of the Council of 9 July 2008 concerning a common framework for merchandising products and the repeal of Council Decision 93/465/EEC, Official Journal of the European Union, 13 August 2008, L-218, pp. 82-128.

verification stage (on delivery) it can be used to show whether the software corresponds to the software that was awarded a type approval.

Blockchain is closely linked with the above. A blockchain is a distributed system of computers that provides its participants with a sort of shared database.
All amendments to this database (the transactions) are recorded, and every participant can check the blockchain's integrity. As long as the majority of the participating computers can be trusted, any attempts to manipulate data that has already been added to the system can be detected. In this way, the use of blockchain could help to boost transparency regarding any modifications made to the software during the transition from design to the final product.

5.3.3 Consideration 3

The European Commission's approach is based on the principle that robotics and AI are still covered by Directive 2006/42/EC. However, at some point in the future, when AI technology delivers machines that are capable of matching human intelligence (termed High Machine Level Intelligence), there will be a shifting of categories. From then on, these 'autonomous robots' may no longer be defined simply as machines. New ethical issues may arise, concerning robot rights, for example (does the robot have the right not to be deactivated?). No further consideration was given to these developments in the report. This is because they involve a much more distant future than the developments in the field of machine learning described in this report.

5.3.4 Consideration 4

This could involve distinguishing between different applications and the extent to which the machine's actions are determined by machine learning algorithms. As machine learning increasingly determines the machine's actions, the machine in question will be shifted into a higher safety category. It might then be suggested that machine learning applications in higher safety categories should be subject to different (or additional) H&S requirements than those governing machine learning applications in lower categories. If these categories are to be at all meaningful, then an inventory of machine learning applications in the workplace will be required.

5.3.5 Consideration 5

We are aware that the Machinery Directive's H&S requirements apply to the manufacturers and developers of machines for both the professional context and the consumer market. Machine learning is expected to feature in an ever wider range of machines in future, including those destined for the consumer market (such as vacuum cleaners and lawn mowers). This touches on the same themes as before, such as the need for transparency regarding machine actions that involve consumers. Accordingly, the H&S requirements presented here are also applicable to machines for the consumer market. It should be noted, however, that the proposed supplementary essential H&S requirements were based purely on research into machine learning applications in industrial machines.

6 References

EESC (2017). Opinion of the European Economic and Social Committee on 'Artificial intelligence – The consequences of artificial intelligence on the (digital) single market, production, consumption, employment and society' *Own-initiative opinion* (2017/C 288/01). https://eur-lex.europa.eu/legal-content/NL/TXT/?uri=CELEX%3A52016IE5369.

Eurofound (2017). Advanced industrial robotics: Taking human-robot collaboration to the next level. Working paper *The Future of Manufacturing in Europe (FOME)* project. https://www.eurofound.europa.eu/sites/default/files/wpfomeef18003.pdf

Eurofound (2018), *Automation, digitalisation and platforms: Implications for work and employment*, Publications Office of the European Union, Luxembourg.

European commission (2018). Commission outlines European approach to artificial intelligence. https://ec.europa.eu/growth/content/commission-outlines-european-approach-artificial-intelligence_en

Future of Life Institute (2017). Asilomar AI principles. https://futureoflife.org/ai-principles/

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts, *Artificial Intelligence,* https://arxiv.org/abs/1705.08807.

Head, M.L., Holman, L., Lanfear, R., Kahn, A.T., & Jennions, M.D. (2015). The extent and consequences of P-Hacking in science. *PLOS BIOL, 13*(3). https://doi.org/10.1371/journal.pbio.1002106

Hofman, J.M., Sharma, A., & Watts, D.J. (2017). Prediction and explanation in social systems, *Science*, *355*, 486-488.

Jansen, A., Van der Beek, D., Cremers, A., Neerincx, M., & Van Middelaar, J. (2018). Opkomende risico's voor arbeidsveiligheid: werken in dezelfde ruimte als een cobot (Emerging risks to occupational safety: working in the same area as a cobot.). TNO report, TNO 2017 R11463.

OSHA (undated). Industrial Robots and Robot System Safety. Section IV: Chapter 4 of OSHA Technical manual. https://www.osha.gov/dts/osta/otm/otm_iv/otm_iv_4.html

Robotics VO (2013). A roadmap for U.S. robotics: From internet to robotics. http://www.roboticscaucus.org/Schedule/2013/20March2013/2013%20Robotics%20Roadmap-rs.pdf

Russell, S. & Norvig, P. (2010). *Artificial Intelligence: A modern approach* (third edition). Pearson.

Samuel, A.L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 3*(3), 210-229.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms.* Cambridge university press.

Silver, D., Schrittwieser, J., Simonyan, K. Antonoglou, I., Huang, A. Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature, 550,* 354– 359.

Steijn, W., Luiijf, E., & Van der Beek, D. (2016). Opkomend risico voor arbeidsveiligheid door inzet van robots op de werkvloer (Emerging occupational safety risks associated with the use of robots in the workplace). TNO report, TNO 2016 R10643.

Steijn, W., Van der Vorm, J., Luiijf, E., Gallis, R., Van der Beek, D. (2016). Opkomende risico's voor arbeidsveiligheid als gevolg van IT-koppelingen van en tussen arbeidsmiddelen (Emerging occupational safety risks associated with IT links from and between work equipment). TNO report, TNO 2016 R10096.

A Appendix: Interview protocol

A.1 Protocol 1. Expert from the field

- Introduction.
- Introduction.
  - Space for interviewees to elaborate on those aspects of their personal background that are particularly relevant to the topic in question.
  - Explanation concerning the practical application.
- What are people's expectations regarding the robots of the future?
  - In terms of mobility.
  - In terms of man-machine interaction.
  - In terms of autonomy.

  - Which functions.
    - Interaction – *explain.*
    - Autonomy - *explain, if necessary emphasize that we are talking about decision-making autonomy here, in which robots determine their own behaviour.*
    - Task complexity – *explain.*
- What will happen if machines are equipped with Machine learning (also known as deep learning) technology?
- What will be the effect of placing increasing numbers of robots in the same environment (or working environment)?
- *Would it be possible to distinguish between the risks of technical failure and those involving human error? There is also human-design, as well as ethical objections/health and safety risks.*

- What are the expected effects?
- What are the expected risks with regard to human safety?
  - *Would it be possible to distinguish between the risks of technical failure and those involving human error? There is also human-design, as well as ethical objections/health and safety risks.*
  - What control measures are there?
- What would change if robots like this were to be used by consumers?

- Legal requirements?
  - Machinery Directives?

- Incidents?

- Is the robot's control system or the central control system protected against external influences (such as viruses or hackers)?
  - If so, how?

- Main reason for choosing a robot.
- Conclusion.

A.2 Protocol 2. Scientific expert

Introduction

- Introduction.
  - Space for interviewees to elaborate on those aspects of their personal background that are particularly relevant to the topic in question.

- Explanation of field of Expertise.
  - Existing applications in robots?
  - Future applications:
    - Short term (within 10 years).
    - Long term (beyond 10 years).
- Risks in the field of occupational safety (as a result of increasing mobility, man-machine interaction, autonomy and machine learning).
  - *Would it be possible to distinguish between the risks of technical failure and those involving human error? There is also human-design, as well as ethical objections/health and safety risks.*
- What would change if robots were to be used by consumers?
- In this connection, what measures to improve safety would be conceivable/necessary?

- To what extent are applications already covered by legislation (such as the Machinery Directive)?
  - Which directives or items of legislation do you use?
  - Have any clashes been identified between elements of the directives used?
  - What is not yet covered by regulations or directives, yet should be?

- Conclusion.

TNO, Healthy Living
Schipholweg 77-89
2316 ZL Leiden
PO Box 3005
2301 DA Leiden
The Netherlands

www.tno.nl

T +31 88 866 90 00
infodesk@tno.nl

Commercial Register number 27376655.