# Ethical Behavior Aspects of Autonomous Intelligent Cyber-Physical Systems

Damien Trentesaux[1] and Stamatis Karnouskos[2]

[1] LAMIH UMR CNRS 8201, Polytechnic University Hauts-de-France, France
damien.trentesaux@uphf.fr
[2] SAP, Walldorf, Germany
karnouskos@ieee.org

**Abstract.** Industry 4.0 fosters the development of more autonomous and intelligent cyber-physical systems that interact with humans in open environments. While the technical aspects are heavily researched, the same does not hold for the accompanying ethical issues that are relevant for the design, development, operation, and maintenance of such systems. In engineering teams and industrialists at large, ethical behavior aspects of autonomous intelligent cyber-physical systems are often seen as an add-on and are rarely appropriately considered during system engineering. The aim of this work is firstly to understand the reluctance of researchers and industrialists and secondly, to justify the need to address ethical considerations in all lifecycle phases of autonomous intelligent cyber-physical systems. A case study in train transportation exemplifies some of the issues discussed, and clearly shows that the development of autonomous intelligent cyber-physical systems without addressing their ethical behavior is incomplete.

**Keywords:** machine ethics · cyber-physical systems · autonomous systems · artificial intelligence · transportation.

## 1 Introduction

Disruptive approaches that emerge from combining the cyber and physical worlds are transforming the industry [34, 35]. Among them, Cyber-Physical Systems (CPS) are seen as the key enabler of the visions behind international efforts to advance industrial systems such as Industrie 4.0 in Germany, “Made in China 2025” in China, “Industrie du Futur” in France or “smart manufacturing” in the USA [11, 37]. Typically, Industry 4.0 is based on several technologies and architectures [30] that are amalgamated from areas such as big data, analytics, autonomous systems, artificial intelligence, simulation, cross-layer integration, cybersecurity, additive manufacturing, and augmented/virtual reality, etc. just to name a few.

On one side, these technologies will enable future CPS to be more intelligent and autonomous. These future CPS will integrate learning abilities to improve their decisions over time. On the other side, future CPS will interact with humans, and they will evolve in a more open environment (e.g., autonomous cars) than today. As a consequence, the impact of their behavior on society (and in a reciprocal way, from society on them as well) must be carefully studied. Such studies are typically related to sustainability, acceptability or ethics. This concerns interdisciplinary works at the frontier of several fields: psychology, sociology, physiology, philosophy, law, and science to name the main ones [27].

This work focuses on the ethical aspects pertinent to the behavior of autonomous and intelligent CPS. It is argued that, even if the theoretical background, named "machine ethics" exists, current research in this field remains scarce, as most of the research efforts focus asymmetrically on the technical aspects of CPS like intelligent sensors and actuators, control loops, safety studies, communication systems, data processing, etc. Although some works deal with the interaction between CPS and humans (human-machine cooperation, human-machine interface, decision support, etc.), there is still need to look specifically to needs and challenges that ethics bring in the industrial CPS [26].

The paper is organized as follows: In section 2, we discuss and characterize the concept of autonomous intelligent CPS. In section 3, the concept of machine ethics is presented. In section 4, we provide and discuss some arguments explaining why researchers and industrialists are reluctant to address ethical behavior aspects of industrial CPS. In section 5, a case study in train transportation is provided as an illustration justifying that machine ethics in CPS must be considered, as otherwise incomplete solutions emerge. Subsequently, in section 6 a critical discussion is carried out and in section 7 a summary of the conclusions are presented.

## 2 Autonomous Intelligent CPS

There are numerous discussions on what Cyber-Physical Systems encompass, and how these are utilized in different domains including Industrial CPS (ICPS) [11], Cyber-Physical Production Systems (CPPS) [6], SmartGrid CPS [9, 16], etc. Overall though we can consider that CPS are systems with embedded software (as part of devices, buildings, means of transport, transport routes, production systems, medical processes, logistics processes, coordination processes and management processes), which record data using sensors, affect physical processes using actuators, evaluate saved data, are connected with other in networks, use globally available data and services and have a series of dedicated human-machine interfaces [1]. A CPS can be viewed as a “System of Systems” as it may be recursively composed of several other constellations of CPS [7, 10]. Complex systems can be formed, however, also commonly federations of CPS may exist and operate in a more coordinated fashion, e.g. a fleet of CPS.

Key aspects of CPS are autonomy and intelligence. For this work, autonomy refers to the degree of freedom a system has regarding potential activities. As an example, the concepts of intelligent agents [20] and holonic approaches [8] are well known for being utilized to realize the autonomy of decision and the autonomy of action. The autonomy of decision refers to the degree of freedom allocated to the system when deciding. For example, it can be associated with

Preprint version of doi:10.1007/978-3-030-27477-1_5

a set of constraints on a search space. The reduction of this degree by choosing one possibility constitutes the act of decision, for example using optimization tools. As mentioned, autonomy concerns also the action led, for example on the real world, through actuators or the digital world through the sending of decisions to apply by others. Assuming at least a certain level of decisional autonomy opens the possibility for a CPS to learn and adapt its decision with time and with its experience and history. Smart agents [19] have widely been used to realize intelligent solutions in different domains, e.g. energy, manufacturing, etc. However, lately, the new re-focus on Artificial Intelligence and its tangible applications is now increasingly integrated into CPS [5] and enables them to learn.

Although certain aspects of a CPS autonomy can be set during the design phase, at the operational stage, these are further refined, especially when considering the interactions of the CPS with other CPS or systems, services, and humans. CPS contain autonomous and cooperative elements and subsystems that interact with each other in different situation-dependent ways [21] and as such patterns and behaviors emerge in larger systems. As a consequence, through the autonomy, CPS inherently support key state of the art technologies that enable them to face demanding dynamic challenges, in terms of flexibility, robustness, adaptation, and reconfigurability, etc. These CPS exhibit self-* properties e.g. are self-adaptive, self-reconfigurable, self-healing, self-organizing [19]. Both CPS behaviors, as well as those emerging from their interactions, are nowadays a result of the application of modern AI approaches and not the result of prescribed rules or hard-coded reactions. Due to the development of autonomous features in modern CPS, that are capable of learning and operating in highly-dynamic situations, we start witnessing more tangible applications of the in real-world that are not confined in controlled environments only e.g. in factory shop-floor, but that can operate in non-deterministic environments e.g. self-driving cars.

A typical example where autonomous intelligent CPS can be seen in action, is that of robots. Robots have been initially developed to ease the repetitive work of humans at the lower costs possible. With time, their design enabled them to realize increasingly complex tasks, in initially highly-controlled environments but nowadays in more flexible ones. Visions position robots to replace the human even in more demanding and common tasks. The interaction with the humans moves to the next level, where the human becomes a supervisor or a collaborator of the robot, and both work jointly towards tasks or goals that can now be realized better and more efficiently. This holds also true for autonomous intelligent CPS; they are intended to improve effectiveness-related indicator, such as completion times and safety levels (assuming that human errors will disappear), while at the same time improve efficiency-related indicators, expressed in terms of cost, energy or environmental impact. Such collaborative efforts also consider other factors that enable a better experience for humans at work, including taking advantage of both e.g. creativity from humans and manual/repetitive work from robots (at the moment).

Preprint version of doi:10.1007/978-3-030-27477-1_5

However, such visions and expectations bear also risks that are not well understood or addressed today. These risks stem from the conjunction of the autonomy of decision and action of autonomous intelligent CPS evolving in an open, uncontrolled environment, as well as their interaction with humans throughout their life cycle, and especially during their use. If such CPS e.g. robots are to operate along with humans, ethical behavior aspects of these autonomous intelligent CPS is an issue that should be addressed.

## 3 Machine Ethics and Intelligent Autonomous CPS

Ethics, as a field of philosophy, engages in concepts of right and wrong. For intelligent autonomous CPS, ethical behaviors become relevant since these are expected to operate within society. However, how actions of the CPS relate to cultural expectations, morality and fairness [22] is something that is still in the early stages. Even for autonomous intelligent CPS which are expected to be introduced in the short term e.g. self-driving cars, ethics must be discussed, investigated or integrated [17].

Two main kinds for ethics from a system engineering point of view are directly relevant i.e. the ethics of the actors involved in the design and production of the CPS and the ethics of the CPS itself during its use [32]. The latter is seen within the area of "machine ethics" [2], which is what is also relevant for this work.

There are several ethical frameworks [13, 28] with various levels of differentiation. From these "deontology" is a normative ethical framework and considers that there are rules that have an absolute quality in them, which means that they cannot be overridden [17]. As such deontology prescribes that one should do what his/her duty is.

The "consequentialism" [15], claims that the right action is one that produces a good consequence among the tested and compared with other possible outcomes. In consequentialist theories the weights given to the outcomes are important, and the outcomes justify the course of action. As an example of consequentialist theory, Utilitarianism is a normative ethical framework that considers as the best action, the one that maximizes a utility function by considering the positive and negative consequences of the choices pertaining to the decision [17]. Other frameworks, of course, exist e.g. relativism, where a range of practices that although considered morally acceptable in some societies, are condemned by others, as well as absolutism which puts forward universally valid moral rules, norms, beliefs, practices, and any deviation or difference is therefore wrong or invalid [17].

One could argue that Asimov's laws of robotics [4] are probably the ones that are more well-known due to the fiction works of the author. However, in practice, if we try to apply the original Asimov's Laws to today's robots the complexities of situations and limits of physical systems as well as the interactions make it challenging [23]. Asimov himself in later works realized similar shortcomings [3]. A revised version of such laws, however, could still be relevant for building guiding meta-ethical frameworks [3].

Preprint version of doi:10.1007/978-3-030-27477-1_5

The fine details of the ethical frameworks, their boundaries, as well as the differentiation within the framework subcategories themselves [28], are not really known or acknowledged in the larger engineering community, which predominantly focuses on the technology itself, and ethics are somehow implied. As ethics, however, are somehow integrated with branches of logic, decision making, engineering, etc., they are increasingly becoming important, as the outcomes of engineering work are now autonomous intelligent CPS which need to take their own decisions and whose actions directly affect humans.

Since ethics is a branch of philosophy, their integration with engineering requires inter-disciplinary interest and at least nowadays, still the study of machine ethics among industrialists and researchers working in control automation, system engineering, and computer sciences is neglected, because of reluctance and misconceptions. The next section is focused on the reasons why one can face such misconceptions and reluctance.

# 4 Ethical Behavior of CPS: Misconceptions and Reluctance

The engineering community that predominantly develops autonomous and intelligent CPS technologies and products, unfortunately, does not feature a good and in-depth understanding of the different angles involved in the area of machine ethics. Here we discuss some empirical views, that have been often expressed when discussing engineering projects relevant to CPS. We take these viewpoints as statements and discuss their origin, context and how to address them.

## 4.1 Viewpoint: "It is someone else's job to do it"

Some of the actors consider that ethics is not part of engineering processes but rather something else. They consider it as a philosophical issue that is challenging but not practically solvable with the tools they know and use. Because philosophical aspects are mostly considered to be relevant to human-related fields, like psychology, law, and sociology, they simply consider that this is someone else's job to investigate once the product is released.

What is different with autonomous intelligent CPS is that an increasing number of them will operate in non-controlled environments, will take real-time decisions based on dynamic situations, and will affect humans directly (e.g. self-driving cars). Understanding in-depth the decision-making process, specifically why a decision was taken, is very challenging in AI-driven CPS [18]. Considering the lack of rule-based systems and the learning empowered by AI, we effectively deal with black-boxes and can not map exactly how a decision is taken and can be affected. New knowledge is highly data-dependent and new data stemming from new experiences (e.g., through the "try and error" mechanism), may alter the existing behavior and fine-tune towards other directions. As an example, a CPS that may be exposed to biased data may be inclined towards new biased behaviors.

Preprint version of doi:10.1007/978-3-030-27477-1_5

The key question here is how to include ethical frameworks in the engineering solutions, that do not exactly prescribe behaviors (this is not seen as feasible nor wished) but still provide operational boundaries within which the CPS can operate. For this to happen, it is clear that ethics cannot be an afterthought, but has to be on equal merit along with other engineering considerations e.g. security and safety, when it comes to design and realize autonomous intelligent CPS.

### 4.2 Viewpoint: "Signing a deontological charter is enough"

Behind this viewpoint, one can find the effort to comply with the adoption of the "Hippocratic oath" in the engineering domain. The idea is that when designers and constructors sign such kind of charter, then they will do their best to design an autonomous intelligent CPS that will behave according to the best interest possible, correctly handle private data, in a secure and anonymous way, etc. More, these systems will host "safety bags" for example limiting their behavior into bounded, controlled situations.

However, such considerations merely diverge from the problem rather than providing a solution to it. Even if CPS designers and manufacturers have paid the maximum attention to design benign CPS, their operational use in non-tested environments or certain dynamic situations means that they can not guarantee their perfect level of reliability. Such situations may arise by chance or as a result of intended misuse e.g. hacking, of a bad maintenance process and plenty of other reasons, and as a result, the CPS may provoke injuries or behave outside the originally envisioned specs. In other words, even if the designers and manufacturers of the CPS have the best intentions and have attempted to integrate guidelines for all scenarios that they consider their CPS might be utilized, there is the chance that dynamic situations may arise, where the CPS will not behave as expected and this might result in biased, unethical or otherwise even dangerous behavior for humans.

Independent of how difficult it is to program ethics into CPS [14], this needs to be realized, and it must adhere to societal and law needs. However, these can not be simply delegated to standalone efforts and must be the result of a discussion coupled with socio-technical aspects and regulation efforts.

### 4.3 Viewpoint: "All you need is a Big Red Button"

The idea to put such a "kill-switch" or "big red button" (BRB) is fostered by international industrialists and some prominent voices including Stephen Hawking, Elon Musk, Bill Gates, Steve Wozniak and others that have publicly warned about the risks of AI and autonomous robots [36]. The idea is to keep under control AI systems [24] e.g. an autonomous intelligent CPS through a BRB that, when pushed, stops them immediately. This idea, although seducing, is effective under very specific conditions and is largely impractical. Such actions are after

Preprint version of doi:10.1007/978-3-030-27477-1_5

all a reaction measure when misbehavior is evident. The aim is to provide psychological assurance to humans, that even if something goes wrong, there will be a possibility to stop the CPS and regain control.

It is questionable if pushing the BRB or kill-switch can be done on time. It also assumes that misbehavior will be detected on-time. But what would be considered as misbehavior? This is a much larger discussion as even cultural differences may lead to classify a CPS behavior as offending or not. Furthermore, even if this is done, what are the effects of such a CPS shutdown? To visualize this, consider that the BRB is pushed for a speeding car on a motorway, a high-speed train traveling at full speed, a plane in the air, etc. Would the human be able to take control of it after the BRB is pushed, and if so, would the remaining time be available for him/her to react to the situation?

BRBs are also a potential misuse item, basically because they shut-down the defenses of a CPS and its intended operation. Assuming that BRBs or similar technologies are used to "teach" the CPS of unwanted behavior, other issues arise. What would the CPS learn from repetitive BRB events? As CPS will not learn as singletons, but as part of a swarm, such events may have unintended effects to a large CPS population and emergent side-effects may appear.

Another question is that even if BRBs are programmed, how long would it take for a CPS to bypass them? Modern AI systems may utilize reinforcement-learning, which aims to always find the optimal operation in its environment. What if the optimal operation is that of finding shortcuts that simply cut out the operator or BRB usage? On the other side of the spectrum, what is wished is a CPS that would self-police, i.e. it would detect its deviating behavior and take measures to either correct it or shut down (push itself the BRB).

### 4.4 Viewpoint: "Don't worry, humans will always be in command"

Some industrialists claim that industry will never sell autonomous intelligent CPS, that cannot be provably controlled by a human supervisor. This approach assumes that CPS operate in a specific context and can always be supervised by the appropriately experienced personnel that will be available and will handle the situation (in situ or remotely). Therefore, the CPS operation will be under constant monitoring of the supervisor and in unexpected hazardous situations s/he will intervene. For such kind of CPS, they see no need to study ethics, as CPS will merely be tools.

Such arguments, however, are very narrow-scoped as they refer to only a very limited number of CPS use cases. This is also a usual fallacy, as it assumes the existence of the "magic human" [31], as shown in Figure 1, who monitors everything in real-time, has perfect knowledge and can take instantly the best decisions (including also a perfect recovery procedure when a BRB is pushed, see the previous section). However such assumptions do not really hold, especially for humans, but also probably not for other advanced AI that could take such a role. Humans can not supervise the different dimensions, and large-scale fleets of the autonomous intelligent CPS envisioned, nor understand their complexity. On

Preprint version of doi:10.1007/978-3-030-27477-1_5

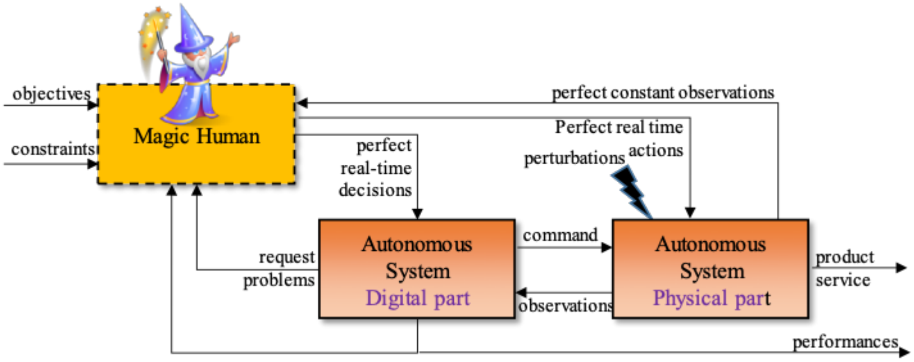
objectives
constraints
Magic Human
perfect constant observations
Perfect real time actions
perturbations
perfect real-time decisions
request
problems
Autonomous System Digital part
command
observations
Autonomous System Physical part
product
service
performances

**Fig. 1.** The fallacy of the "magic human"

the reaction part, humans often overreact or fail to take optimal decisions, which could be dangerous when acting as a remote supervisor facing an emergency.

### 4.5 Viewpoint: "Safety rules and norms do not let any room for ethics"

Autonomous intelligent CPS interacting with humans are considered as critical systems. Because of legal responsibility and accountability issues, they will thus always be governed through safety rules that forbid any degree of freedom for decisions that could lead it in a hazardous state. In CPS, a rule must be thus set for any possible situation to forbid any risky human or CPS behavior. Thus, it may be claimed that incorporating ethics in CPS is not needed, as there is no room for any kind of programmed "common sense" or decisions that the CPS could take according to its programmed ethics. In that line of thought, some may find it difficult if not impossible to program "common sense" or ethics in machines.

Such views, however, are a result of misinformation, limited understanding of technology and its potential spread of applications. Designers and industrialists arguing that everything can be set through rules, clearly have not studied new AI systems and live in the past of "rule-based" approaches that operate in well-mapped environments. Modern systems learn in a non-rule based way, and we actually can not even yet understand how they learn or take the decisions they do [18].

If operational rules are somehow integrated into autonomous intelligent CPS, it is still risky not to identify and accept the available degrees of freedom requiring ethical behaviors to be programmed. It is impossible for every situation to be reliably identified, as that would require full knowledge of the operational environment details and exhaustive testing of all use cases, which is not possible for real-world environments. For example, enabling an autonomous intelligent CPS to adapt in unclassified (unknown) situations an ethical behavior based on the consequential paradigm should be reassuring, at least to ensure that it will

Preprint version of doi:10.1007/978-3-030-27477-1_5

do something to limit the consequences of an accident when evolving in a situation where no rule applies. The results, however, may not be what the designer expects, and what a human would prioritize. A CPS may decide that the best way to extinguish a fire is to remove oxygen from the room, independent if there are still people in it or not. Calculating all consequences of an action is impossible [15], not to mention the weights that need to be put on each consequence, and as such in practice such approaches cannot work. Meanwhile, works must be done to limit as much as possible the risk where no programmed rules apply when ethical issues are at stake.

## 5 Case Study: Train Transportation

To illustrate in practice some aspects of the discussion, a case study, relevant to train transportation is studied. The specific focus is on a dynamic situation, which although hardly deterministic due to the multiple variables in real-world environments, it is probable, and has as the highest goal the well being of humans in critical situations i.e. the triggering of a fire alarm in a train coach. Fires in trains are among the events that present the highest level of risk (human casualties), while at the same time, it is a dynamic phenomenon that is very challenging to fully control in a systematic way. We compare the case where the conductor is a human, with the case of the autonomous train, where the last is seen as an autonomous intelligent CPS transporting people.

### 5.1 The Case of a Human Conductor

Given the high level of risk, train operators have defined procedures for conductors to be followed. In Figure 2, such a procedure stemming from a guide for conductors of an international railway operator, shows the different steps. This procedure is periodically modified to reflect new knowledge from international standards or past accidents. It becomes evident that several ethics-related viewpoints that have been raised in this work, are reflected: the conductor is assumed to be the so-called "magic human", who due to his/her expertise is able to handle whatever happens during the critical situation. For safety and legal issues, the clearly defined steps (rules of the procedure) seem to let no room for the conductor to decide.

The execution of a plan, however, is not as simple as one may assume, and that is mainly due to the dynamic situations that may arise, as well as the conductor's behavior. Reflecting on past emergency situations, it can be seen that some of them were correctly handled by the conductor (who even had to become creative and even come with his own solutions in order to bring passengers to safety), while for some others, the conductor, even following exactly the procedure (e.g., triggering an emergency stop), has made things worse (as s/he ignored additional factors and preferred to stick to the prescribed rules).

As depicted in Figure 2, such an official train operator state-of-the-art procedure contains hidden (un-assumed) consequentialist-based ethical studies (e.g.,

Preprint version of doi:10.1007/978-3-030-27477-1_5

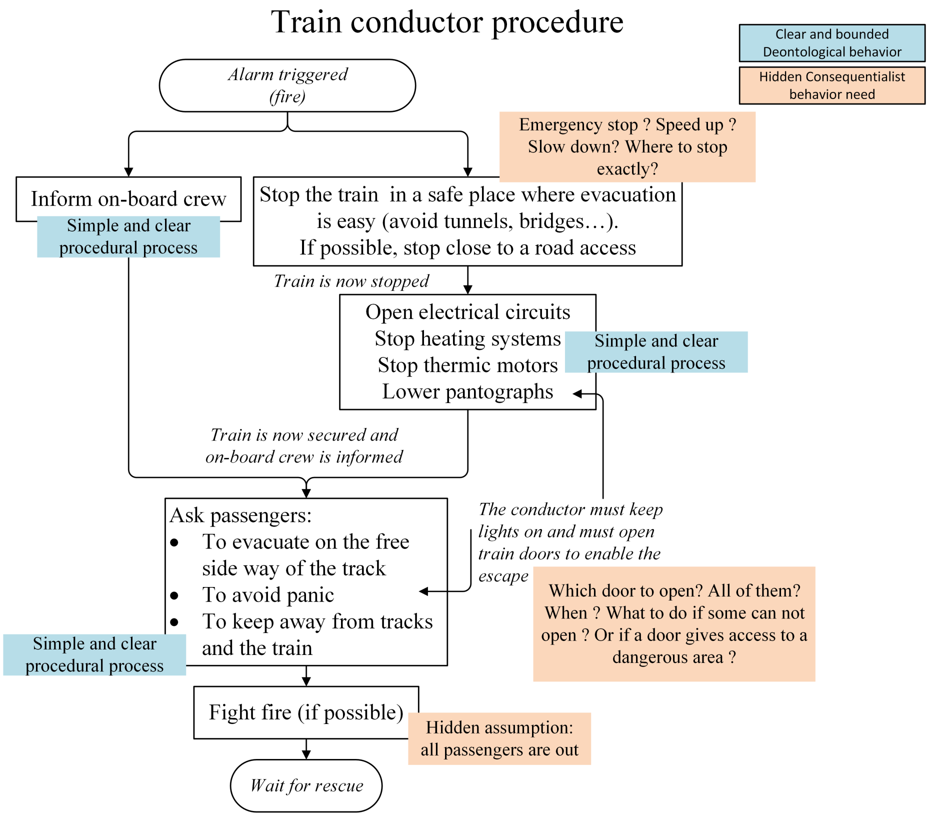
Train conductor procedure
Clear and bounded Deontological behavior
Hidden Consequentialist behavior need
Alarm triggered (fire)
Inform on-board crew
Simple and clear procedural process
Stop the train in a safe place where evacuation is easy (avoid tunnels, bridges…). If possible, stop close to a road access
Emergency stop ? Speed up ? Slow down? Where to stop exactly?
Train is now stopped
Open electrical circuits
Stop heating systems
Stop thermic motors
Lower pantographs
Simple and clear procedural process
Train is now secured and on-board crew is informed
Ask passengers:
To evacuate on the free side way of the track
To avoid panic
To keep away from tracks and the train
Simple and clear procedural process
The conductor must keep lights on and must open train doors to enable the escape
Which door to open? All of them? When ? What to do if some can not open ? Or if a door gives access to a dangerous area ?
Fight fire (if possible)
Hidden assumption: all passengers are out
Wait for rescue

**Fig. 2.** Train conductor regulatory procedure in case of emergency (fire)

"common sense" based decisions) to be made by train conductors. In lack of truly intelligent systems, and due to the existence of rule-based logic, this is common also in other industrial systems e.g. in manufacturing. From our perspective, until AI-based CPS are sufficiently generalizable and can react correctly to critical situations, utilizing the human as an expert is a viable mid-term option. Here CPS will be helping the human with very specific tasks and carry out his decisions.

### 5.2 The Case of the Autonomous Train

The autonomous train is currently under development by major players in the train transportation sector and is expected to become operational in the near-term. Autonomous trains are viewed as autonomous intelligent CPS that interact with humans and infrastructure, while the transportation system overall is seen as a System of CPS. Autonomous trains are intended to be safer (avoiding conductor error as a source of accidents) and energy-aware (optimizing energy

Preprint version of doi:10.1007/978-3-030-27477-1_5

consumption). They also enable the increase of the efficient use of the existing infrastructures by optimizing the available resources and as a result e.g. more passengers per day can be accommodated without additional investment on the infrastructure.

An autonomous train is able to fulfill an assigned complete transportation mission from one railway station to another in an open environment, without human guidance. It must thus be able to perceive, analyze, decide and act in an intelligent and autonomous manner in its operational environment [33]. Current efforts are made in the design of sensing systems and the control of moving operations. Decisions to be taken by the autonomous train seem to be simpler than the ones needed by self-driving cars as e.g. a train evolves on 1D tracks and decisions are mainly related to "stop" and "go" processes. However, they are bundled with more risk and a single error may result in hundreds of casualties due to the large number of passengers they carry at high speeds.

As with any autonomous intelligent CPS interacting with humans, ethical aspects must be considered for the involved processes. In Figure 3 the procedure dealing with an emergency i.e. fire in a train is investigated. Deontological behaviors are concerned with rule-based (simple) tasks while consequentialist ones are advised for more complex ones. This procedure raises several ethical concerns as the autonomous train must decide on how to handle the situation and even in the case of unavoidable casualties take decisions of life and death. Approaching the process with ethical frameworks such as deontological (rule-based decisions) and consequentialist ("what if?" and simulation-based decisions) aspects, may provide some insights on diverging behaviors.

While the exact behavior may be very complex, in this scheme, several actions can be engaged simultaneously by the autonomous train, which may be done sequentially by the human conductor. The task "stop the train ..." requires that the autonomous train simulates different scenarios where the ethical decision to be taken is "where to stop?". This decision would be made using train dynamics models, fire propagation models, and crowd movement models. Typically, a decision minimizing the risk of casualties should be applied. This risk minimization hints a potential utilitarian approach, as the "minimal casualties" need to be calculated. Applying blindly procedural decisions is potentially harmful (e.g., stopping in a tunnel, onto a bridge, close to a motorway, etc.). While simulation may provide some answers to some of these envisioned potential environmental conditions, the dynamic context may complicate things. To illustrate this, should the train on fire attempt to bring passengers near to the next train station where medical help can be provided, if that would imply endangering additional people in train station e.g. if the train explodes or sets other nearby trains on fire?

The same principle applies for the task "open doors". If there is no such simulation, opening all the external doors may create air intakes, accentuating the fire and on the opposite, deciding to close all the internal fire protection doors may lead to suffocation of concerned passengers. The human conductor must have the "common sense" to take care of these aspects, as the autonomous

Preprint version of doi:10.1007/978-3-030-27477-1_5

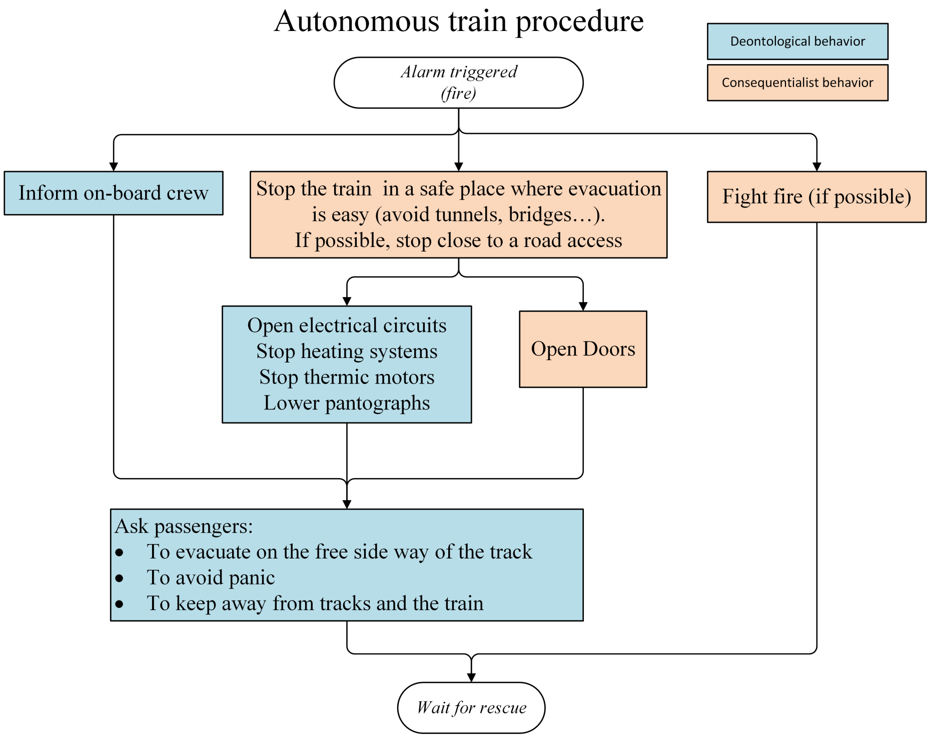
Autonomous train procedure
Deontological behavior
Consequentialist behavior
Alarm triggered (fire)
Inform on-board crew
Stop the train in a safe place where evacuation is easy (avoid tunnels, bridges…). If possible, stop close to a road access
Fight fire (if possible)
Open electrical circuits
Stop heating systems
Stop thermic motors
Lower pantographs
Open Doors
Ask passengers:
To evacuate on the free side way of the track
To avoid panic
To keep away from tracks and the train
Wait for rescue

**Fig. 3.** Autonomous train procedure in case of emergency (fire): a proposal

train will have to. However, we have to be clear that this also may not be optimal, as the conductor may wrongly judge the situation and decide sub-optimal or even wrong actions.

The authors argue that in fact, not only complex tasks but also even "simple" ones may gain from being carefully analyzed, both at the individual as well as at the system level, since applying them could sometimes amplify an ethical issue. For example, opening electrical circuits seems to be such a simple task, which could be handled using a rule-based behavior. Meanwhile, applying it may forbid trapped passengers to breathe fresh air. Thus, from our perspective, each task could gain from being accompanied by a kind of "ethical control system" analyzing constantly the consequences of possible decisions through different balanced ethical frameworks or approaches. However, a trade-off has to be there. The reason is that it is impossible to analyze all decision alternatives, and even then probabilities for each outcome as well as weights will have to be assigned, not to mention that other questions such as how far into the future [12] should consequences be considered e.g. 1h, 1day, 1year?

Preprint version of doi:10.1007/978-3-030-27477-1_5

## 6 Discussion

Taking decisions is a challenging task, especially if one rationally approaches it, attempting to take optimal decisions [12] and not be based on the "common sense" which is subjective and rarely correct. All of the decisions have an additional ethical dimension, which further perplexes the process, especially in critical situations [17].

Humans learn the difference between right and wrong, as they spent their early lives being taught about it from their parents, family, and social circle. However, such a process takes a long time, and although it may be applicable to machines, it is impractical. Machines need to be produced with the capability of distinguishing right from wrong from the begin of their operation.

While some discussions about machine ethics have been going on for quite some time, only recently the advances made in the AI have created results that can be utilized in the real world with tangible applications. Even with such developments at our doorstep, there is insufficient understanding of the ethical aspects involved, not to mention the appropriate education, laws and regulatory frameworks.

The industry is developing autonomous intelligent CPS, and as such, it is only natural that they also make proposals of how ethics can be defined and applied for such emerging products. This is however by no means an easy task as it requires an interdisciplinary approach and it cannot be standalone efforts. After experiences with large industries without oversight, have shown in the past that the results may not always be unbiased e.g. tobacco industry, petroleum industry, etc. What is needed is a pragmatic approach, that will involve industry, but also have sufficient representation from other parts of society.

Setting standards on machine ethics that can be followed, is challenging. Recently IEEE released the second version of its efforts captured in the document titled "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems" [29] which exactly targets this space, and aims towards kick-starting such actions and potential standards. However, moving forward from formulating guidelines towards implementing them in CPS, in a reliable and trustworthy manner means that significant challenges will need to be overcome.

Even then the acceptance of such ethically-behaving large-scale systems is another issue that needs to be addressed. To exemplify this challenge, consider the example of a CPS such as a self-driving car that employs utilitarian ethics [17]. In case of an upcoming accident, such a car may decide that killing its driver is preferable (at society level) than killing five unsuspected pedestrians crossing the street. But if so, who would actually want to buy such a car, especially if this is not a mandated requirement for all cars on the street?

From the ethics perspective, there is no clear answer on what approach should be followed e.g. utilitarian or deontological, etc. as no one-size fits all approach is expected to be found. Among the existing ethical frameworks that can be applied to machine ethics, the pluralism framework is seen as a promising candidate. Pluralism rejects absolutism (that there is only one correct moral truth) and

Preprint version of doi:10.1007/978-3-030-27477-1_5

relativism (that there is no correct moral truth) as unsatisfactory and proposes that there is a plurality of moral truths, arguing that indeed there are universal values (as indicated in absolutism). However, instead of considering that there is only a single set always applicable, it considers that there are many which can be interpreted, understood and applied in diverse contexts (as indicated in ethical relativism) [17]. Therefore, it seems that pluralism covers the requirements that allow the inclusion of diverse cultural aspects, norms and values to be integrated while it can also accommodate globally recognized issues, and approaches.

## 7 Conclusion

Machine ethics are increasingly required as autonomous intelligent CPS are becoming the core of modern infrastructures. However, despite the substantial societal stakes, it is still scarcely sufficiently studied by engineering teams and industrialists working on autonomous intelligent CPS solutions. Driven by some empirical discussions we had with such teams, we have presented some of the common positions raised, and why such arguments are mostly biased or simply bypass the problem rather than addressing it. As it becomes apparent from the discourse in this work, as well as exemplified in the train transportation use-case presented, machine ethics must be addressed on equal merit along with other engineering considerations e.g. security and safety, when it comes to design and realize autonomous intelligent CPS. This also implies that inter-disciplinary research needs to be carried out for researchers to address adequately the new challenges, before the mass introduction of such solutions.

In this paper, the scope was mainly focused on a specific kind of risks, relevant to designers' reluctance and under estimation of the need to address ethical behaviors of future autonomous intelligent CPS. Meanwhile, other kinds of risk exist. For example, the risk to over estimate the ability of Humans to keep emotional and social distance with these future friendly CPS that will behave ethically and pay attention to human welfare [25]. All these risks must be listed and studied urgently.

## Acknowledgement

Surferlab is a joint research lab with Bombardier and Prosyst, which is partially funded by the European Regional Development Fund (ERDF). The authors wish to thank the European Union and the Nord–Pas-de-Calais in France for their support. This research work is also done in the context of the project EFIA "Etude de faisabilité de l'intelligence artificielle pour le train autonome" (feasibility study on artificial intelli-gence for the autonomous train) and "train fret autonome" (autonomous freight train), both funded by the SNCF in collaboration with IRT Railenium research institute. The authors would like to thank the SNCF for its support. Parts of the work presented in this paper are carried out in the context of the HUMANISM No ANR-17-CE10-0009 research program,

Preprint version of doi:10.1007/978-3-030-27477-1_5

funded by the French ANR “Agence Nationale de la Recherche”. This paper contains elements discussed during workshops led within the context of the project “Droit des robots et autres avatars de l’humain” (robots and other human avatar rights) of the IDEX “Université et Cité” of Strasbourg University.

# Bibliography

[1] ACATECH (2011) Cyber-Physical Systems: Driving force for innovation in mobility, health, energy and production. Tech. rep., URL https://goo.gl/Q6WFQN

[2] Allen C, Wallach W, Smit I (2006) Why machine ethics? IEEE Intelligent Systems 21(4):12–17, https://doi.org/10.1109/mis.2006.83

[3] Anderson SL (2007) Asimov's "three laws of robotics" and machine metaethics. AI & SOCIETY 22(4):477–493, https://doi.org/10.1007/s00146-007-0094-5

[4] Asimov I (1950) I, Robot. Bantam Books

[5] Beyerer J, Kühnert C, Niggemann O (eds) (2019) Machine Learning for Cyber Physical Systems. Springer Berlin Heidelberg, https://doi.org/10.1007/978-3-662-58485-9

[6] Biffl S, Lüder A, Gerhard D (eds) (2017) Multi-Disciplinary Engineering for Cyber-Physical Production Systems. Springer International Publishing, https://doi.org/10.1007/978-3-319-56345-9

[7] Cardin O (2019) Classification of cyber-physical production systems applications: Proposition of an analysis framework. Computers in Industry 104:11–21, https://doi.org/10.1016/j.compind.2018.10.002

[8] Cardin O, Derigent W, Trentesaux D (2018) Evolution of holonic control architectures towards industry 4.0: A short overview. IFAC-PapersOnLine 51(11):1243–1248, https://doi.org/10.1016/j.ifacol.2018.08.420

[9] Cintuglu MH, Mohammed OA, Akkaya K, Uluagac AS (2017) A survey on smart grid cyber-physical system testbeds. IEEE Communications Surveys & Tutorials 19(1):446–464, https://doi.org/10.1109/comst.2016.2627399

[10] Colombo AW, Karnouskos S, Bangemann T (2013) A system of systems view on collaborative industrial automation. In: 2013 IEEE International Conference on Industrial Technology (ICIT), IEEE, https://doi.org/10.1109/icit.2013.6505980

[11] Colombo AW, Karnouskos S, Kaynak O, Shi Y, Yin S (2017) Industrial cyberphysical systems: A backbone of the fourth industrial revolution. IEEE Industrial Electronics Magazine 11(1):6–16, https://doi.org/10.1109/mie.2017.2648857

[12] Eisenführ F, Weber M, Langer T (2010) Rational Decision Making. Springer

[13] Ess C (2014) Digital Media Ethics, 2nd edn. Digital Media and Society, Polity Press

[14] Goodall NJ (2016) Can you program ethics into a self-driving car? IEEE Spectrum 53(6):28–58, https://doi.org/10.1109/MSPEC.2016.7473149

[15] Hooker B (2010) Consequentialism. In: The Routledge Companion to Ethics, Routledge, https://doi.org/10.4324/9780203850701.ch37

[16] Karnouskos S (2011) Cyber-physical systems in the smartgrid. In: IEEE 9th International Conference on Industrial Informatics (INDIN), Lisbon, Portugal, https://doi.org/10.1109/indin.2011.6034829

[17] Karnouskos S (2018) Self-driving car acceptance and the role of ethics. IEEE Transactions on Engineering Management pp 1–14, https://doi.org/10.1109/tem.2018.2877307

[18] Lei T, Barzilay R, Jaakkola T (2016) Rationalizing neural predictions. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, https://doi.org/10.18653/v1/d16-1011

[19] Leitao P, Karnouskos S, Ribeiro L, Lee J, Strasser T, Colombo AW (2016) Smart agents in industrial cyber–physical systems. Proceedings of the IEEE 104(5):1086–1101, https://doi.org/10.1109/jproc.2016.2521931

[20] Leitão P, Karnouskos S (eds) (2015) Industrial Agents: Emerging Applications of Software Agents in Industry. Elsevier

[21] Monostori L (2014) Cyber-physical Production Systems: Roots, Expectations and R&D Challenges. Procedia CIRP 17:9–13, https://doi.org/10.1016/j.procir.2014.03.115

[22] Morahan M (2015) Ethics in management. IEEE Engineering Management Review 43(4):23–25, https://doi.org/10.1109/emr.2015.7433683

[23] Murphy R, Woods DD (2009) Beyond Asimov: The Three Laws of Responsible Robotics. IEEE Intelligent Systems 24(4):14–20, https://doi.org/10.1109/MIS.2009.69

[24] Orseau L, Armstrong S (2016) Safely interruptible agents. In: Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence, AUAI Press, Arlington, Virginia, United States, UAI'16, pp 557–566

[25] Pacaux-Lemoine MP, Trentesaux D (2019) Ethical risks of human-machine symbiosis in industry 4.0: insights from the human-machine cooperation approach. In: 14th IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design, and Evaluation of Human-Machine Systems (IFAC HMS)

[26] Pacaux-Lemoine MP, Berdal Q, Enjalbert S, Trentesaux D (2018) Towards human-based industrial cyber-physical systems. In: 2018 IEEE Industrial Cyber-Physical Systems (ICPS), pp 615–620, https://doi.org/10.1109/ICPHYS.2018.8390776

[27] Rault R, Trentesaux D (2018) Artificial intelligence, autonomous systems and robotics: Legal innovations. In: Service Orientation in Holonic and Multi-Agent Manufacturing: Proceedings of SOHOMA 2017, Studies in Computational Intelligence, vol 762, Springer International Publishing, Cham, pp 1–9

[28] Skorupski J (2010) The Routledge Companion to Ethics. Routledge, https://doi.org/10.4324/9780203850701

[29] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2017) Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems. Tech. rep., URL `https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf`

[30] Trappey AJC, Trappey CV, Govindarajan UH, Sun JJ, Chuang AC (2016) A review of technology standards and patent portfolios for enabling cyber-physical systems in advanced manufacturing. IEEE Access 4:7356–7382, https://doi.org/10.1109/access.2016.2619360

Preprint version of doi:10.1007/978-3-030-27477-1_5

[31] Trentesaux D, Millot P (2016) A human-centred design to break the myth of the "magic human" in intelligent manufacturing systems. In: Service Orientation in Holonic and Multi-Agent Manufacturing, Studies in Computational Intelligence, vol 640, Springer International Publishing, pp 103–113

[32] Trentesaux D, Rault R (2017) Designing ethical cyber-physical industrial systems. IFAC-PapersOnLine 50(1):14934–14939, https://doi.org/10.1016/j.ifacol.2017.08.2543

[33] Trentesaux D, Dahyot R, Ouedraogo A, Arenas D, Lefebvre S, Schön W, Lussier B, Cheritel H (2018) The autonomous train. In: 2018 13th Annual Conference on System of Systems Engineering (SoSE), IEEE, https://doi.org/10.1109/sysose.2018.8428771

[34] Tsiatsis V, Karnouskos S, Höller J, Boyle D, Mulligan C (2018) Internet of Things: Technologies and Applications for a New Age of Intelligence. Academic Press, Elsevier

[35] WEC (2015) Industrial Internet of Things: Unleashing the Potential of Connected Products and Services. Tech. rep., URL http://www3.weforum.org/docs/WEFUSA_IndustrialInternet_Report2015.pdf

[36] Winter-Levy S, Trefethen J (2016) Safety first. World Policy Journal 33(1):105–111, https://doi.org/10.1215/07402775-3545918

[37] Xu LD, Xu EL, Li L (2018) Industry 4.0: state of the art and future trends. International Journal of Production Research 56(8):2941–2962, https://doi.org/10.1080/00207543.2018.1444806

Preprint version of doi:10.1007/978-3-030-27477-1_5