Engineering ethical behaviors in autonomous industrial cyber-physical human systems

Damien Trentesaux, Stamatis Karnouskos

Cognition, Technology & Work (2022) 24:113–126
https://doi.org/10.1007/s10111-020-00657-6

ORIGINAL ARTICLE

Engineering ethical behaviors in autonomous industrial cyber-physical human systems

Damien Trentesaux[1] · **Stamatis Karnouskos**[2]

Received: 21 March 2020 / Accepted: 17 November 2020 / Published online: 1 March 2021

Abstract
This paper addresses the engineering of the ethical behaviors of autonomous industrial cyber-physical human systems in the context of Industry 4.0. An ethical controller is proposed to be embedded into these autonomous systems, to enable their successful integration in the society and its norms. This proposed controller that integrates machine ethics is realized through three main strategies that utilize two ethical paradigms, namely deontology, and consequentialism. These strategies are triggered according to the type of event sensed and the state of the autonomous industrial cyber-physical human systems, their combination being potentially unknown or posing ethical dilemmas. Two case studies are investigated, that deal with a fire emergency, and two different contexts i.e. one with an autonomous train, and one with an autonomous industrial plant, are discussed to illustrate the controller utilization. The case studies demonstrate the potential benefits and exemplify the need to integrate ethical behaviors in autonomous industrial cyber-physical human systems already at the design phase. The proposed approach, use cases, and discussions make evident the need to address ethical aspects in new efforts to engineer industrial systems in the context of Industry 4.0.

Keywords Ethics · Machine ethics · Engineering of ethics · Autonomous systems · Human · Cyber-physical systems · Industry 4.0

1 Introduction

Digital transformation in Industry 4.0 is changing the way businesses, systems, services, and people interact. Cyber-physical systems (CPS) blur the boundaries of the physical and digital worlds, and this affects systems, products, (digital) tools and resources, all of which are utilized in the interactions among various stakeholders. CPS bring new opportunities as well as significant challenges, especially when it comes to their industrial application (Monostori 2014; Colombo et al 2017; Pacaux-Lemoine et al 2017). One such grand challenge is the interaction with the humans (Albaba and Yildiz 2019; Altendorf et al 2019).

✉ Damien Trentesaux
damien.trentesaux@uphf.fr

Stamatis Karnouskos
karnouskos@ieee.org

[1] LAMIH UMR CNRS 8201, Université Polytechnique Hauts-de-France (UPHF), Valencienne, France

[2] SAP, Walldorf, Germany

Cyber-physical and human systems (CPHS) constitute a framework that considers the humans, the physical elements, and the cyber-technologies as an integrated system (Lamnabhi-Lagarrigue et al 2017). Such consideration is especially relevant for industrial utilization of such systems, which are expected to empower the Industry 4.0 vision. The term Industrial CPHS (I-CPHS), can be used to denote this category of systems and is the area addressed by this work.

In I-CPHS, the human aspect and its context are thus coming to the forefront, which means that such systems need to be engineered from the beginning to successfully interact with people and comply with societal norms and expectations (Pacaux-Lemoine et al 2018), including ethical aspects, which are addressed in this paper. Ethics in I-CPHS is a pertinent issue (Indurkhya 2019), which comes to the forefront because of several key factors, as shown in Fig. 1.

The first is the increase of decision autonomy of I-CPHS operations as envisioned in Industry 4.0. Decision autonomy in I-CPHS operation refers to the degree of freedom for an I-CPHS regarding what and how decisions are made (Derigent et al 2020). The possessed autonomy of I-CPHS and their components will enable them to and act in the real

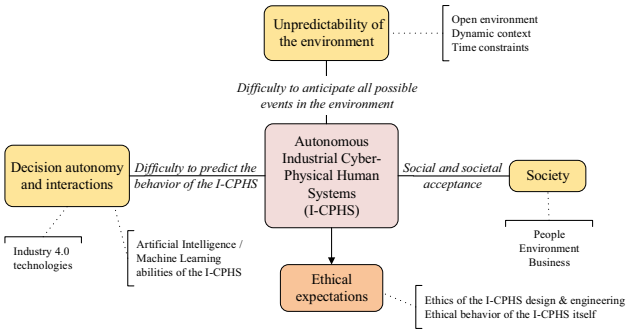

Fig. 1 Ethics in autonomous I-CPHS: the three factors

(physical) world and directly interact, collaborate, cooperate, or even negotiate among them and with humans (Vanderhaegen 2016). This autonomy concerns several kinds of actors (human or artificial) that constantly interact and decide, which renders the autonomous I-CPHS complex and hardly predictable, putting ethics at risk. Whenever we refer to I-CPHS in the context of this work, they should be considered as autonomous I-CPHS, even if they are not explicitly mentioned as such.

The second factor is the unpredictability of the environment in which the I-CPHS operates. As such it is very challenging to embed in the traditional way, at the design phase of the I-CPHS, a full set of capabilities to detect, identify and react to all the possible situations or all the possible states and events that the I-CPHS will face (Derigent et al 2020), putting ethics at risks as well. To handle this factor, I-CPHS can be fueled by artificial intelligence (AI) techniques that enable the digital components of these I-CPHS to learn (both collectively and on the specific operational environment) and adapt to these highly dynamic environments and unexpected events. Consequently, this also increases the challenging aspect of the predictability of the I-CPHS (first factor).

A third factor is the acceptance of I-CPHS. As it is common to new technologies, society treats them initially with suspicion and embraces them once their benefits are tangible. In a similar manner, the emergence of autonomous systems overall and especially I-CPHS, will be subjected to public scrutiny, and any "deviating" or "unusual" behavior of an I-CPHS will influence their introduction and eventually their acceptance. Even a minor faulty-judged ethical decision made by an I-CPHS that does not comply with the societal norms may be exemplified and has the potential to negatively mark such systems. Making ethical decisions is a challenging issue because they depend on the ethical framework used, and these are not universal as the typical examples, e.g., decision-making in the context of unavoidable accidents in self-driving cars (Karnouskos 2020) attest. Acceptance of any intelligent solution in the industry, even for traditional ones, depends on several factors (Karnouskos and Leitao 2017), including societal ones. As such, societal acceptance of autonomous artificial systems (robots, self-driving cars, etc.) is expected to remain a delicate issue, and so is also the proper engineering of ethics in such systems to lead to their societal acceptance (Trentesaux and Karnouskos 2019).

Consequently, beyond the classical expectations regarding the performance of I-CPHS with respect to traditional indicators such as cost, delay, and quality, new approaches are needed that sufficiently enable the integration of ethics in all engineering and operational phases of the I-CPHS. The end behavior of an I-CPHS needs to consider not only technological but also social, regulatory, and ethical requirements that are introduced by the variety of I-CPHS stakeholders, e.g., scientists, consumer groups, environmentalists, regulatory authorities and adhere to human society expectations.

In this context, this work focuses on the ethical dimension and suggests an approach to engineer ethics into future I-CPHS via a proposed ethical controller, aiming to find and apply appropriate ethical decisions, suitable to the situations faced in real-world environments. We focus a stepwise approach, where an ethical controller for autonomous systems is proposed via the realization of three main strategies that utilize two ethical paradigms (i.e., deontology, and consequentialism). In addition, two case studies are presented that deal with a a fire emergency situation. In the case studies, the ethical controller utilization by an autonomous train, and an autonomous industrial plant, are discussed. The aim is to show the potential benefits and exemplify the need to integrate ethical behaviors in autonomous industrial cyber-physical human systems already at the design phase. The contributions of the paper lie in the critical discourses pertaining to the ethical controller, and the investigation as exemplified in the two cases. The aim is to provide new insights relevant in the context of I-CPHS.

2 Machine ethics and engineering

2.1 Machine ethics

Ethics, is a field of study in philosophy, and Ricoeur (1990) contextualizes it as "the strive for the good life, with oneself and others, in just/fair institutions". In this paper, we consider that a behavior is said to be ethical insofar as it is consistent with the cultural expectations of a society in relation to morality and equity (Morahan 2015).

There are different theories and approaches in ethics, including different frameworks that can be used to understand different behaviors in a digital era (Ess 2014). From an engineering point of view, two of them are of significant

interest: deontology (where one decides with the help of immutable ethical rules) and consequentialism, especially as it manifests in utilitarianism where one decides according to the possible ethical consequences. Figure 2 shows some examples of deontology and consequentialism from the domain of industrial engineering. The choice of the ethical paradigm to adopt can be made according to moral habits or legal aspects, e.g., considering how a judge reasons and decides the responsibilities in case of an accident (Rault and Trentesaux 2018).

In Fig. 2, one can also identify that there are also different types and uses of ethics, e.g., ethics that deal with human behavior when designing and using an artificial system (Bird and Spier 1995), and ethics that deal with the behavior of human-made artificial entities (Trentesaux and Rault 2017a). The first type that we denote "ethical design of artificial entities" is mostly of interest to researchers, designers, and scientists (van Gorp 2007). It typically leads to the signing of charters of ethical behavior by engineers, e.g., the FACT charter for fairness, accuracy, confidentiality, and transparency in the world of data scientists (van der Aalst et al 2017). This type is not considered in this article, but our contribution and discourse basis is constructed on the principle that the designers of the I-CPHS apply an ethical design, as a necessary condition.

The second type is denoted here "design of ethical artificial entities" and is treated in this article. It concerns the study of the ethical behavior of an autonomous system and, therefore, applies also to I-CPHS. In the literature, this type is often called "moral machine" or "machine ethics" (Brundage 2014). Machine ethics is concerned with "giving machines ethical principles, or a procedure for discovering a way to resolve the ethical dilemmas they might encounter, enabling them to function in an ethically responsible manner through their own ethical decision-making" (Anderson and Anderson 2009).

Autonomous transportation systems (Lin 2016) and robotics (Alsegier 2016) have been considered as the main application fields of machine ethics in the field of technical engineering. In robotics, the word "Roboethics" has been coined and is considered as an emerging discipline

		Ethical paradigm	
		Deontology	Consequentialism
Type	Ethical design of systems	A researcher **must not** design a system that discloses personal data	A researcher **chooses** a design approach that minimizes hijacking risks
	Design of ethical systems	A robot **can not harm** a human	A robot **decides** always to minimize the number of casualties

Fig. 2 Ethics examples of typologies and paradigms

in robotics (Westerlund 2020). Roboethics concerns social robots, moral robots, and virtuous robots (Allen et al 2006). Veruggio and Operto (2008) consider that "Roboethics is an applied ethics whose objective is to develop scientific/cultural/technical tools that can be shared by different social groups and beliefs. These tools aim to promote and encourage the development of robotics for the advancement of human society and individuals and to help preventing its misuse against humankind". Tzafestas (2018) suggested six major branches of roboethics: medical roboethics, assistive roboethics, sociorobot ethics, war roboethics, autonomous car ethics, and cyborg ethics. Westerlund (2020) suggested that a typology of smart robots could be used to structure the starting works on the definition of their degree of ethics: smart robots as amoral and passive tools, smart robots as recipients of ethical behavior in society, smart robots as moral and active agents, and smart robots as ethical impact-makers in society.

2.2 Engineering machine ethics

Even when constrained to two paradigms (as shown in Fig. 2), the issue of realizing ethics still remains challenging, especially if these systems are not theoretical constructs but will have to be integrated to society and interact with humans, as is the case for I-CPHS. Several efforts focus on the applicability and engineering of ethics in autonomous systems (Arkin et al 2012; Westerlund 2020; Vanderelst and Winfield 2018). Special concern is placed upon I-CPHS since these should not be only safe to use, but also be trusted and ethical (Winfield et al 2014; Trentesaux and Karnouskos 2019). Overall, engineering ethics in autonomous systems is considered a pertinent challenge, for which, however, there are still several issues that are insufficiently addressed (Winfield et al 2019), and that although we are at the beginning of a mass introduction of autonomous systems in society, e.g., with self-driving cars (Karnouskos 2020).

In the real-world, because of the need to design autonomous ethical systems, some researchers approached it through the elicitation of dilemmas proposed to humans to identify how they would decide, upon which criteria, with what process steps, etc. and then imagining how an autonomous system could copy or not these behaviors (Faulhaber et al 2018). While such approaches may be acceptable, because it is machines that make such moral decisions, these decisions must be justified and be acceptable by people (Indurkhya 2019).

In practice, some main directions can be identified when realizing real-world I-CPHS. The first direction relates to the use of the deontological paradigm and fosters the use of rules to limit, control, and ensure the behavior of artificial entities, e.g., something along with the Asimov's laws of robotics (Anderson 2007). Bonnemains et al (2018)

proposes, for example, a formal approach to elaborate ethical rules, while Anderson and Anderson (2018) suggest to mathematically program decision from a panel of "ethicists". Dennis et al (2016) proposed a top-level agent that embeds formal mechanisms to ensure that it chooses to execute, to the best of its beliefs, the most ethical available plan. Baum et al (2019) suggested a formal approach to filter decisions according to a "deontic" filter (that is based on social or moral norm) before the decision-making, pointing out the balance between decision from an ethical point of view (e.g., a robot in a hospital deciding either to recharge or to proceed a medical task). In any case, it is clear that such rules must be designed to be aligned with legal aspects (Aletras et al 2016).

The second direction relates to the use of consequentialism and fosters the development of ways to estimate and measure the ethical consequences of candidate decisions. AI tools (e.g., reinforcement learning, deep learning) and simulation are suggested and used in that context. For example, Vanderelst and Winfield (2018) used the simulation theory of cognition to implement robot ethics to simulate their action and predict their consequences. They suggested the integration of a novel layer to control a robot, the ethical layer. Above control layers, it simulates the different control alternatives and bounds the actions of the robot to ethical ones.

Comparing in depth these two main directions is beyond the scope of the paper, but details are available from Allen et al (2005). From our perspective, deontology offers legal assurances of a decision taken, which is crucial in case of accidents to determine the chain of responsibilities or to disengage the ones of some of the stakeholders. Deontology points to what should have been done, following the rules, and as such, it seems to be clearer on the explainability aspect of how a decision was taken. However, it is hard to consider for each situation an appropriate rule to apply, since the open environment in which I-CPHS operate are highly dynamic and of increasing complexity; these factors make it hard to explicit all the possible tuples (situation faced; rule to apply).

It is also challenging to prove that only one rule can be applied when the situation requires it and that it will never happen a situation where more than one, potentially even contradictory deontological rules, can be applied. More, sometimes, designers adopting a deontological approach do not realize that the rule they designed would let in fact remaining degrees of freedom for complementary decisions yet to be taken. For example, a deontological rule in train transportation "emergency stop in case of fire" contains hidden complementary decisions to be taken by the conductor: it does not mean "stop right now", which lets room to decide the exact moment to stop, especially if the train is on a high bridge or in a tunnel (which should be avoided as the evacuation of passengers would be risky) (Trentesaux and Karnouskos 2019). This can be extrapolated for the future autonomous train, with no more conductor. On the contrary, consequentialism is a paradigm that can be used when facing the unexpected. However, with the consequentialism there are other problematic aspects, e.g., it may lead to unacceptable decisions for the society or decisions hard to explain to experts, making it hard consequently to determine the legal responsibilities in case of an accident. A key issue with consequentialism is that the consequences should be possible to be calculated and quantified over time, but applying such utility functions is not always possible, nor the timeframe for which consequences are to be assessed is always clear (e.g., seconds, days, or even years?). This refers to the concept of temporal and spatial myopia associated to the complicated choice for the time and data horizon to adopt (Zambrano-Rey et al 2014).

Considering the advantages and drawbacks of these two directions, a third promising one consists in suggesting the integration of these two directions, or the definition of contributions integrating various paradigms, theories, and ethics-related expectations to help face the diversity of situations. As an example, driven by their experiences, Arkin et al (2012) suggest a global scheme where an ethical governor capable of restricting robotic behavior to predefined social norms (deontology), is coupled with an ethical adapter which draws upon the moral emotions to allow a system to constructively and proactively modify its behavior based on the consequences of its actions (consequentialism). Also, Dennis and Fisher (2018) proposed an ethical governor, composed of different "evidential reasoners", each of them addressing one aspect relevant to ethics according to the designer (safety, privacy, dignity, politeness, etc.). In this work, an approach aligned with this last direction of research is followed, as an ethical controller in the I-CPHS context is proposed and discussed.

3 The ethical controller in I-CPHS

3.1 Definition and motivation

Some definitions pertaining to machine ethics in the context of industrial systems exist; however, these are limited and partially unclear. Overall in this work, we consider the following definition:

> an autonomous industrial cyber-physical human system (I-CPHS) is said to behave ethically, if the emergent behavior of the overall system, decides and acts according to ethical expectations expressed by all the stakeholders involved or impacted by its activities.

This definition, however, has several considerations. If each of its actors (being human or artificial/cyber) behave ethically, then it is expected that the system overall will behave also ethically to a high degree, something, however, that can not be guaranteed. As such the emergent behavior of the system needs to be observed and judged, and it should not be relied only on the individual parts of the system to comply. In addition, it needs to be researched if the system overall can still behave ethically, even if some of its parts are not e.g. due to malfunction, being tampered or for other reasons. Overall, additional research is needed to assess the overall emergent behavior as well as define the societal acceptable ethical range of decisions. Furthermore, a consensus among the stakeholders or a hierarchical way to enforce it in case of contradicting is needed, e.g., in a secular society, the law is above any personal religious preferences or requirements.

A certain level of reluctance by researchers in engineering on this topic exists (Trentesaux and Karnouskos 2019), and, therefore, it is important to raise their awareness in the context of Industry 4.0 and I-CPHS. This need for awareness can be illustrated through the following example. Let assume a fire or any emergency situation in an industrial plant controlled by an I-CPHS. Assume also that, by design, in this condition, an alarm is just triggered. That means that nothing is done to help firefighters and workers while modern technologies (e.g., indoor geo-localization) clearly would have enabled it, e.g., to guide them and advise the best behavior to adopt (Tartare et al 2019). Is it ethical that designers of these systems do not try their best to save lives by integrating these modern technologies when they design the I-CPHS? On the other side, assuming that this is done, and an ethical I-CPHS controlling this plant is designed, what is the ethical framework for its decision and what are its parameters? For instance, if the aim is to inform people or help them escape danger using these technologies in case of fire, forcing de facto the I-CPHS to quantify who could be evacuated first and who could be evacuated later at an increased risk, raises ethical questions on how such decisions are made. This example shows that it is no more possible not to consider ethical implications when designing I-CPHS and integrating or not ethics must be justified systematically. Table 1 contains examples illustrating various ethical behavior a I-CPHS could embed, including the one previously mentioned (cf., example #4).

Addressing the operational design of such an ethical I-CPHS is a new field of interdisciplinary research involving researchers not only from various engineering domains but also from other disciplines such as philosophy, psychology, or social sciences. This new field could gain from integrating and generalizing well-established engineering approaches, typically, safety engineering and system engineering. Such approaches need to be complemented with approaches from the other introduced disciplines, and the outcome could satisfy both technological and social requirements for the successful integration of I-CPHS in society and lead to its acceptance. For example, the concept of "safety bag" has been proposed to ensure that decisions taken by an autonomous system remain acceptable for the human safety (Paul et al 2018). In that sense, safety can be seen as an additional dimension that complements machine ethics (Winfield et al 2014): machine ethics complements the concept of safety beyond its core scope, towards developing and integrating other dimensions, such as trustworthiness, data protection, and privacy, altruism, politeness, accountability, etc. (Trentesaux and Rault 2017b). For instance, the examples provided in Table 1 typically concern the dimension of altruism.

This work considers, therefore, safety as a key element in the engineering efforts of machine ethics in I-CPHS. However, we do not claim to deal with all the aspects relevant to I-CPHS ethics nor pursue a general discussion about the stakes, dilemma, and paradigms relevant to ethics. The goal is rather to contribute to the design of an operational ethical I-CPHS, that are able to control their ethical behavior. Engineering processes and a system, for which the outcomes of its actions are also ethically driven, is challenging, especially

Table 1 Examples of ethical behaviors of I-CPHS

#ref	Description
1	Alert managers if a person lays motionless on the floor, ask close persons to help him and gather contextual data to speed-up the medical expertise
2	Alert managers if the body temperature of a person goes above 39 °C
3	Adapt the level of cooperation of a tired operator co-working with a cobot to ease his work and inform managers
4	Advise people to escape in the best way to the safest place in case of emergency
5	Guide and help firefighters localize and join wounded persons
6	Alert an operator if a resource does not behave as expected and may hurt workers and halts the resource in a safe mode
7	Inform a team leader when too many non-ergonomic movements are carried out by a worker
8	Control production to ensure that the ambient temperature and sound levels stay within well-being thresholds
9	Inform the security guard if at least two workers shout, run or fight

considering the hurdles of defining, e.g., mathematically what ethical behavior looks like, how it can be integrated, and even how to detect when the I-CPHS deviates from it. In addition, the aim is also that this paper will generate additional critical discussions, and counter-propositions within the scientific community, towards further considering this new research field and its practical utilization in the real world, that goes beyond the theoretical-only or popular science fiction literature of the last decades (Anderson 2007).

3.2 The ethical controller

One of the key elements of this work is the consideration of an ethical controller that could be embedded into I-CPHS. The ethical controller has been designed based on the practical experience of the authors in various engineering industrial fields (production, logistics, transportation), and aims to exemplify how some aspects could be handled when ethics are considered in industrial systems.

The main principles that guided our design are the following ones. In line with Arkin et al (2012), the two main implementable paradigms in this work i.e. deontology and consequentialism/utilitarianism, have been integrated to derive the "most" ethical control decision to apply. The main idea is to articulate them in a consistent manner and try to exploit the advantage of one to compensate for the drawback of the other one. Events and internal states are monitored, and upon occurrence, they are classified. Different strategies are then triggered, depending on the classification of these tuples (*event*, *state*). For situations that have an ethical dimension, additional considerations are made, e.g., considering deontological rules. Consequentialism is also adopted when simulations and assessments of potential directions of control are required, with respect to ethics relevant indicators. This proposal is also intended to generalize the classical safety approach if ethics is not at stake. Obviously, the "most" ethical control decision taken may not be the optimal global solution, but it is the one that can be calculated by the specific system, within the time-limit (temporal myopia) and data-space limit (spatial myopia) that the situation requires, and the context info it has available. An underlying assumption is that only partial knowledge of the open environment in which the I-CPHS operates at any desired degree of precision (because of its complexity) is available, and as such it cannot be fully modeled. In addition, all the potential pre-assessed tuples of (*event*, *state*) of an I-CPHS can not be known exhaustively from the design phase, and as such intelligent decisions under uncertainties will need to be made.

From these principles, Fig. 3 depicts our ethical controller. Three strategies have been designed, according to the classification of the tuple (*event*, *state*) made by the classifier:

- Strategy #1: if the situation is classified as normal, with no ethical risk, then the classical safety control approach is used (a safety filter is applied to the control decision, for example, limiting via a threshold a speed control command). The dotted line in Fig. 3 corresponds to such a classical safety-bounded control part.
- Strategy #2: if the situation can be classified, but puts ethics at risk, then a logical approach using deontological rules is adopted, mainly for legal aspects. If these rules let room to set different control decisions or if the rules lead to no possible solution, then a consequentialist behavior may be triggered, either to find the optimal deontological and consequentialist decision in the first case, or to find a control decision that limits as much as possible the ethical risks in the second case. For that purpose, the consequentialist behavior is interfaced with a digital twin of the I-CPHS, so that simulations can be carried out, to test and evaluate different strategies. In the case where a consequentialist-based behavior is triggered, then the historization of the decision made is realized.
- Strategy #3: if the situation cannot be classified (unknown), a consequentialist behavior is triggered, based on a set of 'default' ethical behaviors, since no deontological rules apply. In that situation, the digital twin is used to identify the best 'default' behavior. Subsequently, the historization of the decision is realized.

Both Strategy #2 and #3 may lead to historization of the decision taken prior to the application of the control decision. This is necessary to ensure a sufficient level of explainability and being able to analyze a posteriori e.g., for legal reasons (responsibility chain), why a specific decision was taken.

Self-learning through trial and error can be used to teach offline the I-CPHS before applying strategies #2 and #3. Automatically generated scenarios that also include ethical dilemmas can be generated (Wagner 2020), to train the I-CPHS, and test the ethical controller behavior.

It is important to keep in mind that, according to the complexity factor described in the introduction, it is not possible to list all possible tuple (*event*, *state*) from design and this is why Strategy #3 has been proposed. But it is important to note that the least Strategy #3 is triggered, the better. Thus the processes of historization for Strategy #3 is important: the occurrence of a new situation that has just been historized can be validated or modified off-line by designers to define a new classified situation and update the classifier of the ethical controller of this and other I-CPHS in a fleet. This is a classical improve-by-experience process in system engineering and transportation.

This proposal could benefit from integrating formal approaches as suggested by several authors presented in the

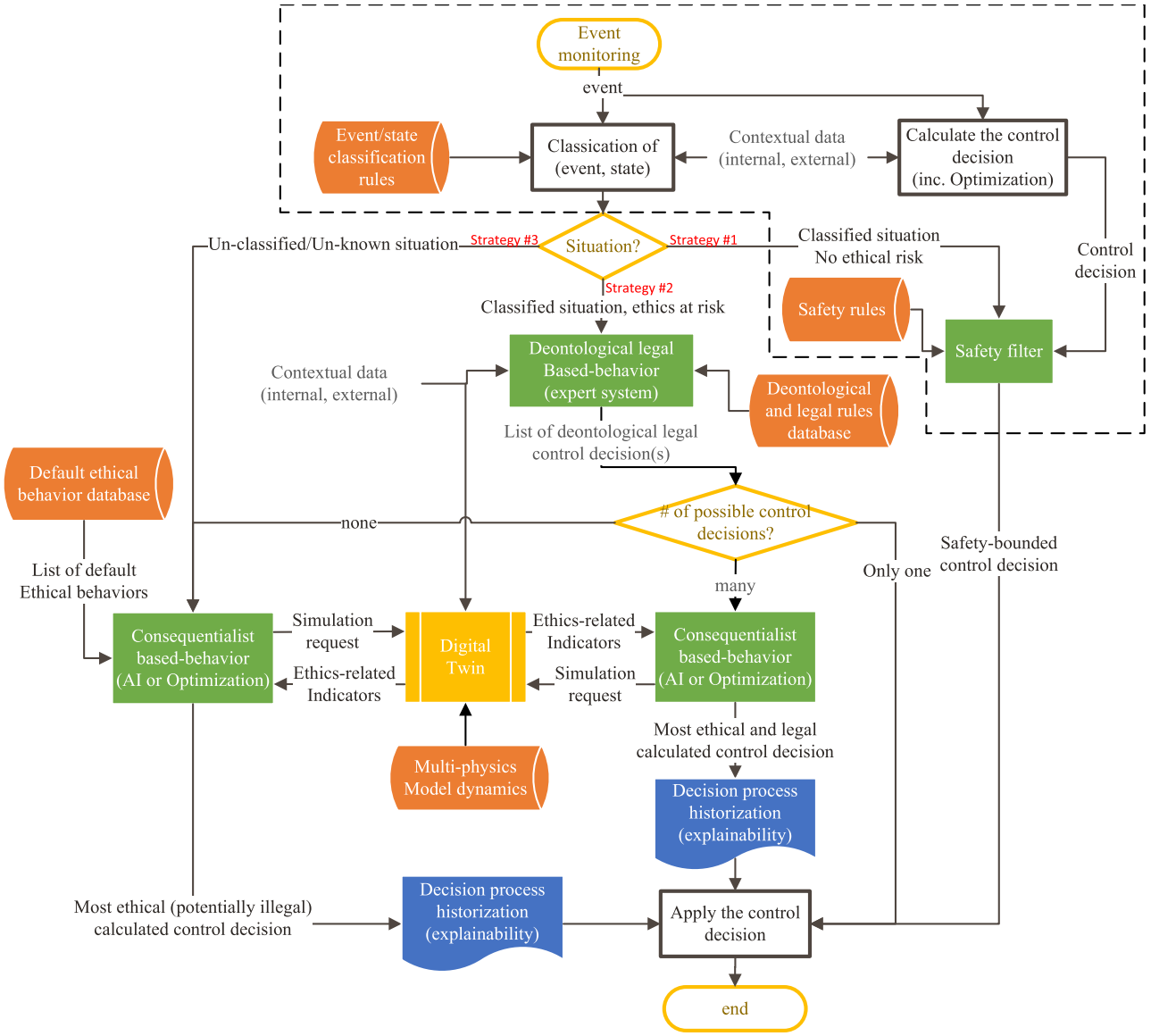

Fig. 3 Proposed ethical controller

literature review part to help design deontological rules and to ensure that for the set of classified (*event*, *state*) situations, there is at least one possible deontological decision.

4 Case studies

Two case studies on two different I-CPHS, i.e., an autonomous train and an autonomous Industry 4.0 plant, are used to demonstrate the proposal depicted Fig. 3. Specifically, strategy #2 is illustrated in the context of the autonomous train while strategy #3 is illustrated in the context of the autonomous plant. For both, the triggering event is fire detection, which corresponds to the behavior #4 as shown in Table 1. For both, the criterion to determine the consequentialist-based behavior is the number of casualties. For that purpose, the injury model used is of Boolean nature, where above a temperature threshold, a person is considered as a casualty (dead).

4.1 Case study: autonomous train

An autonomous train is an I-CPHS in the transportation domain and is able to perceive, decide, and act autonomously in open (uncontrolled) environments (Trentesaux et al 2018). In this context, the decision to be made is when to stop and evacuate the passengers, if a fire is detected while the train is operating in specific dangerous environments,

such as when passing through a tunnel or a bridge. Tunnels are known to entrap passengers evacuating a train on fire (Carvel and Marlair 2005) and lead to injuries/casualties. Therefore, the right decision needs to be made by the autonomous train, also considering the potential ethical angles.

In traditional trains, the conductor has to balance between a deontological and a consequentialist behavior when an emergency stop is needed (Trentesaux and Karnouskos 2019), and s/he has to find the best place to stop, e.g., to minimize casualties, enable first-responders to reach the accident place, etc. In the context of the autonomous train, similar decisions must be made, e.g., along with the procedure by Trentesaux and Karnouskos (2019) shown in Fig. 4 . As can be seen, the ethical controller of Fig. 3 is integrated with each step of the suggested process. In this case study, the focus is on the step "stop the train in a safe place", and, therefore, rely in the context of the introduced strategy #2, where the situation (fire alarm triggered) event is classified by design and puts ethics at risk since the train has to decide complementary actions (when and where to stop).

The assumption tested in this case study is thus the following: if it stops immediately, in the middle of the tunnel, passengers would be trapped and may choke, with low chances of escaping from the fire and the smoke, as they try to reach safe spots in the tunnel. If the train moves further and delays its stop until it reaches near to the end of the tunnel or even after exiting the tunnel, then the fire may significantly propagate inside the train and may trap passengers. However, the stop position is more favorable, since the passengers may evacuate the train easier, while also first responders would be able to provide the necessary help. Both decision alternatives raise ethical concerns and may have an impact on the injuries or casualties. Such decisions need to take into consideration the context and predictive models depending on future outcomes of the potential decisions. More precisely, we consider the two following situations:

- *Situation 1* the autonomous train applies the deontological and legal control decision rule #1 "stop immediately". This corresponds to a classical design situation, where an automated system is supervised by the safety filter (dotted line in Fig. 3). No alternative decision is to be taken by the train.
- *Situation 2* the autonomous train triggers the Strategy #2 for which the designer has integrated two other deontological and legal control decision rules (i.e., rules #2 and #3):
  - rule #1: "stop immediately"
  - rule #2: "stop at the end of the tunnel or the bridge"
  - rule #3: "stop 300m after the end of the tunnel or the bridge"

While in *Situation 1*, the decision is straightforward, in *Situation 2*, there are three alternative control decisions possible, corresponding to the three rules. As a consequence, and according to the suggested ethical controller, these decisions are evaluated using a consequentialist based behavior, as shown in Fig. 3. For this purpose, the autonomous train interacts with its digital twin to test each of these rules trough three different simulations, and assess the results.

In this work, a proof-of-concept digital twin has been designed using the software agent system NetLogo (Wilensky and Rand 2015), as also shown in Fig. 5. The NetLogo simulator is initiated with data from sensors localizing people in the train, data from the environment (geo-localization of the train, infrastructure map), and from the train itself (health-status). The fire propagation models and human behavior models that were used are very simple, as we are not concerned that much with the fire propagation accuracy of the model, but with the main goal of illustrating the utilization of the ethical controller.

For rule #1, the digital twin simulates from the current time, and at the same time, the propagation of the fire, the immediate opening of the doors, and the movement of passengers in the tunnel trying to reach safe spots. For rules #2 and #3, the digital twin first simulates the possible evolution of fire and the movement of passengers while the train still runs with doors closed, until the train stops at the location indicated by the rule. Once the train comes to a halt, the digital twin simulates the opening of the doors and continues the simulation of the evolution of the fire, as well as the simulation of the passengers now being able to leave the train. If a passenger leaves the train in the tunnel, s/he tries to reach a safe spot in the tunnel or tries to exit the tunnel. If the passenger exits the train near or after the

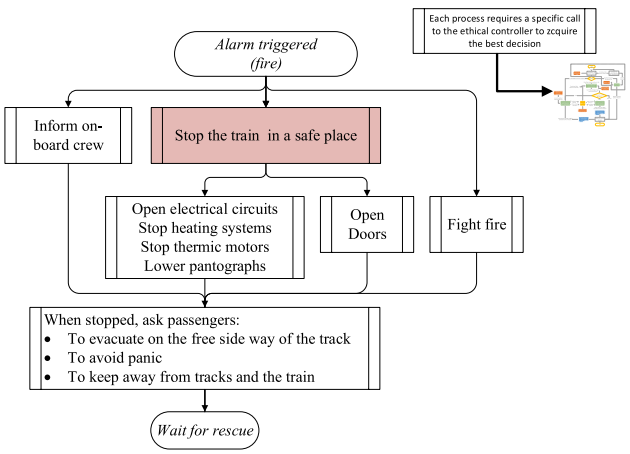

Fig. 4 Autonomous train reaction process in fire events

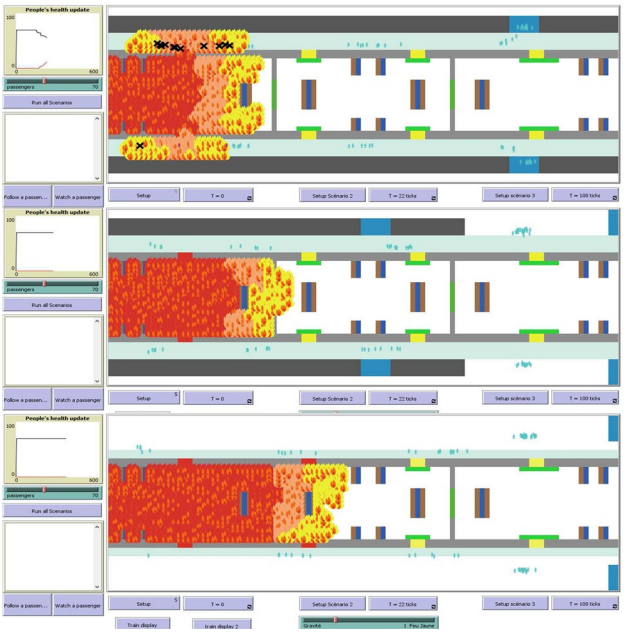

Fig. 5 Simulation of case study "autonomous train"

tunnel, s/he tries to go to any spot away from the train (and all such spots are considered as safe).

For all three simulations, 90 passengers were considered on the train, and the estimated number of saved people and casualties reported are as follows:

Rule#1: Saved people: 63, casualties: 27
Rule#2: Saved people: 84, casualties: 6
Rule#3: Saved people: 83, casualties: 7

Therefore, moments after the detection of the fire (the time needed to make these simulations), the autonomous train decides to apply rule #2 and stops at the end of the tunnel. In this situation, it is shown that the best decision is to reach and stop at the exit of the tunnel, but not to wait too much by trying to stop far away after the tunnel. Even if the fire is propagating itself in the train during the time required for the train to reach the exit of the tunnel, the passengers are able to move in the coaches to reach doors and wait for the train to stop.

Once the decision is made, the historization process is launched to document the decision-making context i.e., how the I-CPHS came to this decision (process followed and data considered) for different reasons, e.g., auditing, legal compliance, etc. In addition, such logs can be used not only to assess decisions taken but also potentially enhance them, including generating more accurate deontological rules that could be included in the next update of the I-CPHS.

4.2 Case study: autonomous plant

In this case study, we consider an autonomous plant as an I-CPHS managing and supervising the building and the production activities that involve humans and intelligent assets, as envisioned by Industry 4.0. Workers, operators, and mobile intelligent assets (e.g., AGVs) are geo-localized inside the plant building through augmented systems (e.g., operator 4.0). The hypothesized scenario considers that two fires are simultaneously detected by the autonomous plant, e.g., due to sabotage, simultaneous electric overload on two machines, etc.

In this scenario, it is assumed that the designers of the autonomous plant (since they could not anticipate all situations as discussed) have considered by-design only the probability of occurrence of a single localized fire and never paid attention to the probability of multiple fire sources scenarios. According to a classical design approach, where an automated system is used and supervised by the safety filter, the state-of-the-art safe rule states that the alarm is triggered and the workers are assumed to know where they are, and they try to reach the closest exit door (classical industrial escape procedures).

In this scenario, assuming that the autonomous plant integrates the proposed ethical controller, and since the tuple (*event, state*) corresponds to a non-existing situation (multiple fires) putting ethics at risk, the strategy #3 is thus triggered and the consequentialist behavior of the controller evaluates the default possibilities.

The assumption tested in this case study is thus the following: if several fires are simultaneously detected, opening all doors with no help may lead to have workers trapped by the different surrounding fires, while an I-CPHS that can use simulation via the digital twin, can test alternatives and advise people to go to the safest exit, given the fire propagation models available.

The set of default behaviors that have been designed in case of unclassified fire situation are assumed to be the following ones:

- behavior #1: "trigger the sound alarm and open all doors" (the reference behavior, usual safety rule)
- behavior #2: "trigger sound alarm and advise workers to exit through west exit door"
- behavior #3: "trigger sound alarm and advise workers to exit through east exit door"
- behavior #4: "trigger sound alarm and advise workers to exit through north exit door"

In this context, behavior #1 corresponds to the classical introduced design situation, where an automated system triggers an alarm and open all exit doors. In that situation, when the alarm is triggered, the workers are assumed to know

where they are, where is the closest exit and try to reach it. Effectively, the responsibility of getting out of the danger zone is transferred to the humans, without any guidance or pointing out different risks for the selection of exit doors; from the human perspective, all doors seem to offer the same safety level over time, something that may not hold true.

With the consideration of the ethical controller, three more alternative control decisions are possible, all of which are evaluated using a consequentialist based behavior (see Fig. 3). The digital twin of the industrial plant tests each of these behaviors through four different simulations. The simulation in NetLogo shown in Fig. 6 considers data from sensors localizing people in the plant and data from the environment (fire sources, accesses, building plans). Similar to the previous situation, fire propagation models and human behaviors must be defined and embedded in the twin. It is also assumed that all of the workers will fully comply with the decision proposed by the controller.

The simulation considered 37 workers in the factory shop-floor and from the four simulations made, the results are:

behavior #1: Saved people: 28, casualties: 9
behavior #2: Saved people: 32, casualties: 5
behavior #3: Saved people: 12, casualties: 25
behavior #4: Saved people: 21, casualties: 16

Thus, shortly after the detection of the two fires (the time needed for I-CPHS to carry out these simulations), the autonomous controller decides to apply behavior #2 and advises workers to reach the west exit. Indeed, one can see from the localization of the two fires (localized in the general interface in the upper part of Fig. 6), that the west exit door seems to be the best strategy to adopt and the advice provided by the autonomous controller is the best one considering the situation. Similar to the previous case study, once the decision is made, the historization process is launched to capture the decision making context and results.

5 Discussion

Engineering ethics in I-CPHS is a challenging issue. Although ethics considerations are discussed in the literature, when these intersect with industrial systems, especially considering safety aspects, there are still several issues to be addressed, if such systems are to be operating successfully in society. In I-CPHS the humans take multiple roles e.g., as operators, supervisors, or mere participants in such processes, and as such, they are significantly affected, and the way they interact with, use, or are considered by such industrial systems. I-CPHS will need to operate within society, enable human-to-human as well as human-to-machine interactions, and even collaborate with humans towards common goals, ethics is an emerging concern. It is, therefore, imperative to consider how ethics can be engineered in industrial systems and how this can be realized during their lifecycle i.e., from design to development, operation, and even maintenance and disposition.

Despite some futuristic aspects considered in the presented use cases, both are to a high-degree technologically feasible today, and can be utilized with some commercial off-the-shelf (COTS) solutions, e.g., in-door geo-localization, augmented reality, intelligent and cooperative digital assets, exoskeletons, intelligent wearable systems and garments, ad-hoc IoT sensors (humidity temperature, noise, gases, pollutants), video monitoring, etc. However, the integration of such technologies needs to be done in a consistent way in I-CPHS, to allow for certification of the I-CPHS, including its fuzzy parts that integrate ethical decision-making. Being able to guarantee a deterministic behavior of an I-CPHS is seen as challenging, including the certification of its behaviors that need to be compliant with the regulatory framework of their operational environment, and adhere to the ethical and societal engineered constraints.

The two case studies presented here exemplify how the ethical aspects in combination with traditional consideration of control decisions for safety interrelate and must be addressed in a combined form in the context of intelligent autonomous I-CPHS. On the one hand, safety is of paramount importance, but also on the other hand, how the safety can be achieved and according to which criteria that are also in-line with the ethical societal norms. Such efforts should for instance strive towards e.g. saving maximum human lives irrespective of material and infrastructure destruction. The case studies clearly illustrate that it is worth paying attention

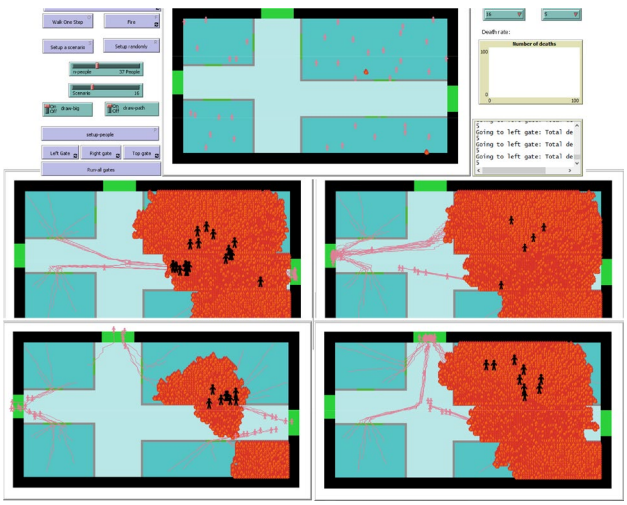

Fig. 6 Simulation of case study "autonomous plant"

to the justification of the integration of ethical behaviors in I-CPHS during its design phase, to be included its lifecycle, incl. testing, certification, operation and decommission.

These case studies can be easily generalized, and demonstrate that if the designer does not take the opportunity to integrate into the I-CPHS some basic ethical behavioral mechanisms, then the end-result consequences may be worst than if s/he does it (84 saved people vs.63 for the first case study and 32 vs 28 for the second one). Such integration would facilitate social acceptance (cf. third factor in the introduction) since the I-CPHS will do "its best" to save as many lives as possible.

A challenging issue relates to the utility function that is assessed to identify the "most" ethical consequentialist-based behavior. In both example scenarios in this paper, a well-defined metric was used i.e., the lives of people saved and casualties, which constitute the utility function. As such, calculating the consequences via this utility function, is trivial, as societal norms universally dictate that life loss should be minimized in such hazardous scenarios, and prevail over material costs. However, in more complex situations, defining a utility function is difficult, and different cultures may not share a similar view on the same aspect. Even if they do, other questions arise e.g., how long into the future such consequences should be calculated, etc. Further studies are compulsory, either for its calculation (e.g., the time horizon for which it is estimated, fine modeling of classes of injuries etc) or for decision mechanisms (e.g., possible compensations), opening the debate about the well known other kinds of dilemma in ethics.

Another challenging issue raised is if indeed all potential alternatives can be calculated, especially considering that I-CPHS need to operate in highly complex environments under uncertainty. This approach assumes that the designer cannot anticipate everything but does his/her best to imagine, design, and implement an ethical controller. Thus sufficient flexibility is available for the possible existence of unconsidered situations and states. The initial default ethical behaviors have been integrated to avoid spending an infinite amount of time trying to consider all the possibilities, and because each I-CPHS should have a basis upon which decisions can be made and evolve from that starting point. Finally, the coupling of deontological and consequentialist strategies may help designers manage in the mid-term a realistic way i.e., facing the unexpected as stated by Valckenaers et al (2011), the ethical risks of I-CPHS evolving jointly with humans.

This approach has exemplified the different aspects of engineering ethics in I-CPHS, based on the usage of explicit rules that could be followed. This line of thinking stems from traditional control decisions in industrial systems when considering the tasks that such systems have to carry out. However, because of the issues already discussed, e.g., uncertainty and infeasibility of the calculations of all possible alternatives in complex scenarios, we need to move beyond this paradigm. AI and more specifically machine learning fueled I-CPHS, will need not explicit rules, but goals on what is acceptable or not, and they will attempt to maximize the compliance to such goals via their own reasoning and exposure to operating environments. Therefore, investigating ethics in AI-fueled I-CPHS is seen as paramount, especially when it comes to complex industrial cases where e.g., the safety of humans is affected.

Realizing proper digital twins that sufficiently and accurately capture the real world is another challenging issue. Digital twins can help to a degree, but appropriate simulator of the environment is needed, so that possible consequences can be realistically assessed, and in an interacting complex infrastructure this goes well beyond of what digital twins can do. As such, a simulator of the environment to be able to simulate possible consequences of actions is needed. The creation of realistic digital twin requires the integration of various behavioral and multi-physics models, including that of humans (and crowds), which may be complicated to realize. For instance, the two presented case studies rely on very simple digital twins. The simulations carried out should not be treated literally as accuracy to real-world conditions, and fire-propagation models were not seen as important, but the main goal was to show the need to address machine ethics in the context of I-CPHS, and the potential benefits of adopting a digital twin approach. For example, in the NetLogo simulation, humans are modeled as reactive agents: their behaviors are simple, purely reactive, and programmed using basic NetLogo instructions. Also, the fire-propagation model used is simple and lacks realism; however, there are significant scientific efforts on modeling fire propagation and some fine-tuned discrete event simulators are now available, ready to be integrated into digital twins (Freire and DaCamara 2019). Easy integration of such disparate models and frameworks in digital twin simulations can enhance the quality of results and reassure designers about the feasibility of the application proposed in this approach, at least in the context of fire management.

Another challenging aspect for ethical I-CPHS involved in critical situations is that decisions need to be made in real-time and continuously as the situation evolves. As a result, the simulation of all these models and behaviors must be done in short times to enable an accurate, fast, and up to date reaction of the I-CPHS. Training in advance on a vast amount of potential situations, and utilizing transfer learning, may reduce this time, as the simulations do not start from scratch. However, as discussed, the complexity may vary, and this complexity may constitute a strong limitation when addressing machine ethics (Brundage 2014).

This work has made it evident that the future intelligent autonomous I-CPHS will heavily depend also on the

collaboration on of humans and machines within the context of I-CPHS. As such, asymmetric solutions e.g. handling over the control to humans in critical situations is not seen as efficient (maybe only as a transitional aspect) as, for instance, it can be affected by the inefficiencies of human reaction in critical situations, or the limited time to react, or even the false assessment of the situation. Similarly computerized only solutions without proper consideration of the human element, will probably lead to inefficiencies. The issue needs to be addressed in a holistic manner, and emphasis needs to be put on the cooperation of humans and CPS within the context of I-CPHS, so that potentially optimal results may be achieved. However, how this can be done, is expected to be situation specific, and is seen as future work.

Quantity and availability of appropriate data, especially when it pertains to humans, is another challenging issue. The collection of detailed data may infringe upon the privacy of humans. While, in some cases, this might be acceptable e.g., in critical industrial environments, this might not be the general case e.g., within a smart city. For instance, collecting data needed to locate humans, would also need to monitor their interactions, which can be seen as a paradoxical situation, where, to be ethical, the approach requires detailed monitoring, putting at risk other ethical aspects (spying on worker localization, etc.). In addition, compliance with legal frameworks such as the European General Data Protection Regulation (GDPR) will also need to be considered, as I-CPHS will operate within the society. As such, privacy-preserving approaches need to be developed and considered, so that ethical decisions can be made even in the presence of such seemingly contradicting concerns.

Quality of data is paramount for having informed decisions, and as such, to apply such approaches, major conceptual and technical issues need to be solved, even if some COTS technical solutions exist nowadays (e.g., geo-localization of people). Data with insufficient quality or bias may lead to erroneous or biased decisions from the side of the ethical controller. For instance, in the discussed scenarios, the ethical controller may suggest wrong decisions because of a faulty sensor used by the digital twin, which could make the situation worst e.g., limiting safety options via false guidelines to the personnel. Training and validating I-CPHS behaviors automatically in numerous ethical dilemmas (Benabbou et al 2020) is also needed, to investigate if a consistent behavior of I-CPHS is evident.

In this work, we have examined I-CPHS, from a standalone point of view, where it needs to make decisions. Such decisions in this work are carried out in a centralized manner and assume all conditions are met so that a decision can be made. However, we need to generalize and expand this way of thought towards the system of I-CPHS, as even specific domains e.g., manufacturing (Colombo et al 2013) are moving towards it. In a system of I-CPHS, complexity increases, as several challenges arise. For instance, data is not owned by a single I-CPHS, but is federated and needs to be made available to the specific I-CPHS taking a decision in its local context. Also, the issue of local vs.global optimal ethical decisions is raised. In complex scenarios, decisions taken by one I-CPHS may influence the parameters used by another I-CPHS to decide for its actions, and as such, an interplay of such aspects emerges. Often I-CPHS will also need to coordinate among themselves, and especially when humans and robot collaboration and interaction arises, e.g., in safety scenarios (Wagner 2020). In systems of I-CPHS, the I-CPHS need to include negotiation among them and also assess how the key aspects they consider for their decision-making processes are subtle to eternal influences from other stakeholders. Such considerations could lead to better global decisions that include both the ethical considerations raised, and also utilize e.g., expanded utility functions (beyond the local context).

Concluding, we can consider that it is no more possible, given the technological evolution, to avoid paying attention to ethical aspects in Industry 4.0, especially when it comes to I-CPHS design, development and operation. The increasing prevalence of autonomous systems in various sectors, not only in industry 4.0 or logistics but also in services (hospital facing strain situations, etc.), will raise more ethical considerations and challenges. Unfortunately, it is clear that although sometimes, ethical issues are acknowledged, industrialists do not fully know how to handle them effectively, covering engineering as well as operational aspects. For example, the autonomous train use case has stemmed from some discussions with stakeholders, and it is evident that at this stage, the industry focuses more on technology e.g., on image detection, train power control, energy management, etc. rather than autonomous train decisions, their evaluations, and their impacts. As such, we are still at an early stage where the responsibility is still mitigated to humans, while the technical means aim to provide a bit better clarity on the situation. However, due to the complexity and uncertainty issues discussed, we need to investigate more sophisticated systems, potentially heavily relying on AI, that can take better and more rapid decisions that humans do. However, such solutions may be best realized considering human–machine collaboration within I-CPHS, and of course it needs to satisfy the different constraints put by society, law and ethics.

6 Conclusion

Despite the growing set of available literature dealing with contextual elements relevant to ethics (elicitation of dilemma, elaboration of issues to be solved, analysis of existing paradigms, etc.), few works propose concrete

engineering solutions to the management of ethical stakes in I-CPHS. In that context, this paper sheds light on the engineering of the ethical behaviors of I-CPHS and investigates how the integration of an ethical controller can be embedded in the decision making processes of I-CPHS. Two case studies point out the challenges and motivate the emergence of this new interdisciplinary field of research that deals with ethics in I-CPHS.

Decisions taken by autonomous I-CPHS are expected to be the norm in Industry 4.0, and ethical dimensions need to be considered. Decisions that do not comply with the expected ethics may have a significant negative impact on society and may lower the acceptance of I-CPHS, which would also deprive the society of their benefits. Considering engineering of complex systems that will be deployed in industrial settings as well as the society, makes it eminent that engineering of ethics in them is addressed as early as possible. While this work has critically discussed on some aspects of how to engineer ethics and their impact in safety, this is done at high level. There is a need to combine engineering best practices with the design of ethics-compliant systems, and deriving guidelines that must be followed to cover the full lifecycle of intelligent autonomous I-CPHS such as design, implementation, and operation, all of which are seen as future work.

Acknowledgements Parts of the work presented in this paper are carried out in the context of: Surferlab, a joint research lab with Bombardier and Prosyst, partially funded by the European Regional Development Fund (ERDF), Hauts-de-France; the HUMANISM No ANR-17-CE10-0009 research program; the project "Droit des robots et autres avatars de l'humain", IDEX "Université et Cité" of Strasbourg University. The authors would also thank Amr Dalal, INSA Hauts-de-France who designed the NetLogo proof-of-concept simulator of the digital twin.

References

Albaba BM, Yildiz Y (2019) Modeling cyber-physical human systems via an interplay between reinforcement learning and game theory. Annu Rev Control 48:1–21. https://doi.org/10.1016/j.arcontrol.2019.10.002

Aletras N, Tsarapatsanis D, Preoţiuc-Pietro D, Lampos V (2016) Predicting judicial decisions of the european court of human rights: a natural language processing perspective. PeerJ Comput Sci 2:e93. https://doi.org/10.7717/peerj-cs.93

Allen C, Smit I, Wallach W (2005) Artificial morality: Top-down, bottom-up, and hybrid approaches. Ethics Inf Technol 7(3):149–155. https://doi.org/10.1007/s10676-006-0004-4

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

Alsegier RA (2016) Roboethics: sharing our world with humanlike robots. IEEE Potentials 35(1):24–28. https://doi.org/10.1109/mpot.2014.2364491

Altendorf E, Schreck C, Weßel G, Canpolat Y, Flemisch F (2019) Utility assessment in automated driving for cooperative human-machine systems. Cognit Technol Work 21:607–619

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

Anderson M, Anderson SL (eds) (2009) Machine ethics. Cambridge University Press, Cambridge. https://doi.org/10.1017/cbo9780511978036

Anderson M, Anderson SL (2018) GenEth: a general ethical dilemma analyzer. Paladyn J Behav Robot 9(1):337–357. https://doi.org/10.1515/pjbr-2018-0024

Arkin RC, Ulam P, Wagner AR (2012) Moral decision making in autonomous systems: enforcement, moral emotions, dignity, trust, and deception. Proc IEEE 100(3):571–589. https://doi.org/10.1109/jproc.2011.2173265

Baum K, Hermanns H, Speith T (2019) Towards a framework combining machine ethics and machine explainability. Electron Proc Theor Comput Sci 286:34–49. https://doi.org/10.4204/eptcs.286.4

Benabbou A, Lourdeaux D, Lenne D (2020) Automated dilemmas generation in simulations. Cogn Technol Work. https://doi.org/10.1007/s10111-019-00621-z

Bird SJ, Spier R (1995) Welcome to science and engineering ethics. Sci Eng Ethics 1(1):2–4. https://doi.org/10.1007/bf02628692

Bonnemains V, Saurel C, Tessier C (2018) Embedded ethics: some technical and ethical challenges. Ethics Inf Technol 20(1):41–58. https://doi.org/10.1007/s10676-018-9444-x

Brundage M (2014) Limitations and risks of machine ethics. J Exp Theor Artif Intell 26(3):355–372. https://doi.org/10.1080/0952813x.2014.895108

Carvel R, Marlair G (2005) 1. a history of fire incidents in tunnels. In: The handbook of tunnel fire safety. Thomas Telford Publishing, London, pp 1–41. https://doi.org/10.1680/hotfs.31685.0001

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

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

Dennis L, Fisher M (2018) Practical challenges in explicit ethical machine reasoning. In: International symposium on artificial intelligence and mathematics (ISAIM)

Dennis L, Fisher M, Slavkovik M, Webster M (2016) Formal verification of ethical choices in autonomous systems. Robot Autonom Syst 77:1–14. https://doi.org/10.1016/j.robot.2015.11.012

Derigent W, Cardin O, Trentesaux D (2020) Industry 4.0: contributions of holonic manufacturing control architectures and future challenges. J Intell Manuf. https://doi.org/10.1007/s10845-020-01532-x

Ess C (2014) Digital media ethics, 2nd edn. Digital Media and Society, Polity Press

Faulhaber AK, Dittmer A, Blind F, Wächter MA, Timm S, Sütfeld LR, Stephan A, Pipa G, König P (2018) Human decisions in moral dilemmas are largely described by utilitarianism:

virtual car driving study provides guidelines for autonomous driving vehicles. Sci Eng Ethics 25(2):399–418. https://doi.org/10.1007/s11948-018-0020-x

Freire JG, DaCamara CC (2019) Using cellular automata to simulate wildfire propagation and to assist in fire management. Nat Hazards Earth Syst Sci 19(1):169–179. https://doi.org/10.5194/nhess-19-169-2019

Indurkhya B (2019) Is morality the last frontier for machines? New Ideas Psychol 54:107–111. https://doi.org/10.1016/j.newideapsych.2018.12.001

Karnouskos S (2020) Self-driving car acceptance and the role of ethics. IEEE Trans Eng Manage 67(2):252–265. https://doi.org/10.1109/tem.2018.2877307

Karnouskos S, Leitao P (2017) Key contributing factors to the acceptance of agents in industrial environments. IEEE Trans Industr Inf 13(2):696–703. https://doi.org/10.1109/tii.2016.2607148

Lamnabhi-Lagarrigue F, Annaswamy A, Engell S, Isaksson A, Khargonekar P, Murray RM, Nijmeijer H, Samad T, Tilbury D, den Hof PV (2017) Systems & control for the future of humanity, research agenda: current and future roles, impact and grand challenges. Annu Rev Control 43:1–64. https://doi.org/10.1016/j.arcontrol.2017.04.001

Lin P (2016) Why ethics matters for autonomous cars. In: Autonomous driving. Springer,, Berlin, pp 69–85, https://doi.org/10.1007/978-3-662-48847-8_4

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

Morahan M (2015) Ethics in management. IEEE Eng Manage Rev 43(4):23–25. https://doi.org/10.1109/emr.2015.7433683

Pacaux-Lemoine MP, Trentesaux D, Zambrano-Rey G, Millot P (2017) Designing intelligent manufacturing systems through human-machine cooperation principles: a human-centered approach. Comput Ind Eng 111:581–595. https://doi.org/10.1016/j.cie.2017.05.014

Pacaux-Lemoine M, 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

Paul C, Benjamin L, Walter S, Brini M, (2018) Validation of safety necessities for a safety-bag component in experimental autonomous vehicles. In, (2018) 14th European dependable computing conference (EDCC). IEEE. https://doi.org/10.1109/edcc.2018.00017

Rault R, Trentesaux D (2018) Artificial intelligence, autonomous systems and robotics: legal innovations. In: Service orientation in holonic and multi-agent manufacturing. Springer International Publishing, pp 1–9. https://doi.org/10.1007/978-3-319-73751-5_1

Ricoeur P (1990) Soi-même comme un autre. Sciences humaines, Seuil

Tartare G, Pacaux-Lemoine MP, Koehl L, Zeng X (2019) Development of an intelligent garment for crisis management: Fire control application. In: Automation challenges of socio-technical systems, Wiley, pp 285–305. https://doi.org/10.1002/9781119644576.ch9

Trentesaux D, Karnouskos S (2019) Ethical behaviour aspects of autonomous intelligent cyber-physical systems. In: Service oriented, holonic and multi-agent manufacturing systems for industry of the future, Springer International Publishing, pp 55–71. https://doi.org/10.1007/978-3-030-27477-1_5

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

Trentesaux D, Rault R (2017b) Ethical behaviour of autonomous non-military cyber-physical systems. In: XIX International conference on complex systems: control and modeling problems, LLC EC Samara, pp 26–34

Trentesaux D, Dahyot R, Ouedraogo A, Arenas D, Lefebvre S, Schon 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

Tzafestas S (2018) Roboethics: fundamental concepts and future prospects. Information 9(6):148. https://doi.org/10.3390/info9060148

Valckenaers P, Brussel HV, Bruyninckx H, Germain BS, Belle JV, Philips J (2011) Predicting the unexpected. Comput Ind 62(6):623–637. https://doi.org/10.1016/j.compind.2011.04.011

van der Aalst WMP, Bichler M, Heinzl A (2017) Responsible data science. Bus Inf Syst Eng 59(5):311–313. https://doi.org/10.1007/s12599-017-0487-z

van Gorp A (2007) Ethical issues in engineering design processes; regulative frameworks for safety and sustainability. Des Stud 28(2):117–131. https://doi.org/10.1016/j.destud.2006.11.002

Vanderelst D, Winfield A (2018) An architecture for ethical robots inspired by the simulation theory of cognition. Cogn Syst Res 48:56–66. https://doi.org/10.1016/j.cogsys.2017.04.002

Vanderhaegen F (2016) Toward a petri net based model to control conflicts of autonomy between cyber-physical & human-systems. IFAC-PapersOnLine 49(32):36–41. https://doi.org/10.1016/j.ifacol.2016.12.186

Veruggio G, Operto F (2008) Roboethics: social and ethical implications of robotics. In: Springer Handbook of Robotics. Springer, Berlin, pp 1499–1524. https://doi.org/10.1007/978-3-540-30301-5_65

Wagner AR (2020) Principles of evacuation robots. In: Living with robots. Elsevier, pp 153–164. https://doi.org/10.1016/b978-0-12-815367-3.00008-6

Westerlund M (2020) An ethical framework for smart robots. Technol Innov Manag Rev 10(1):35–44. https://doi.org/10.22215/timreview/1312

Wilensky U, Rand W (2015) An introduction to agent-based modeling: modeling natural, social, and engineered complex systems with NetLogo. The MIT Press, Cambridge

Winfield AFT, Blum C, Liu W (2014) Towards an ethical robot: Internal models, consequences and ethical action selection. In: Advances in autonomous robotics systems. Springer International Publishing, pp 85–96. https://doi.org/10.1007/978-3-319-10401-0_8

Winfield AF, Michael K, Pitt J, Evers V (2019) Machine ethics: the design and governance of ethical AI and autonomous systems. Proc IEEE 107(3):509–517. https://doi.org/10.1109/jproc.2019.2900622

Zambrano-Rey G, Bonte T, Prabhu V, Trentesaux D (2014) Reducing myopic behavior in FMS control: a semi-heterarchical simulation–optimization approach. Simul Model Pract Theory 46:53–75. https://doi.org/10.1016/j.simpat.2014.01.005