Towards Minimally Conscious Cyber-Physical Systems

A Design Philosophy

Jiří Wiedermann

Jan van Leeuwen

Technical Report UU-PCS-2020-02
July 2020

Center for Philosophy of Computer Science
Department of Information and Computing Sciences
Utrecht University, Utrecht, The Netherlands
www.cs.uu.nl

Series: UU-PCS

Department of Information and Computing Sciences
Utrecht University
Princetonplein 5
3584 CC Utrecht
The Netherlands

Towards Minimally Conscious Cyber-Physical Systems A Design Philosophy*

Jiří Wiedermann[1] and Jan van Leeuwen[2]

[1] Institute of Computer Science of Czech Academy of Sciences and Karel Čapek Center for Values in Science and Technology, Prague, Czech Republic
jiri.wiedermann@cs.cas.cz
[2] Dept of Information and Computing Sciences, Utrecht University, the Netherlands
J.vanLeeuwen1@uu.nl

Abstract. Incidents like the crash of Lion Air Flight 610 in 2018 challenge the design of reliable and secure cyber-physical systems that operate in the real-world and must cope with unpredictable external phenomena and error-prone technology. We argue that their design needs to guarantee *minimal machine consciousness*, expressing that these systems must operate with full awareness of (the state of) their components and the environment. The concept emerged from our recent efforts to develop a computational model for conscious behavior in robots. It leads to a full-fledged design philosophy for 'cognitive' cyber-physical and cyber-physical human systems, based on analogies with the design of weak AI's. Making systems minimal machine conscious leads to more trustworthy systems, as it strengthens their behavioral flexibility in varying environments and their resilience to operation or cooperation failures of their components or as a whole. The notion of minimal machine consciousness has the potential to become one of the defining attributes of Industry 4.0.

> *We need to break down the concept of consciousness into different aspects, all of which tend to occur together in humans, but can occur independently, or some subset of these can occur on its own in an artificial intelligence. [...] We can imagine building something that has some aspects of consciousness and lacks others.*
>
> Murray Shanahan [5], 2018

Keywords: cyber-physical systems, design philosophy, digital twins, Industry 4.0, minimal machine consciousness, reliability, safety, self-control.

1 Introduction

Aircraft crashes like that of Lion Air Flight 610 and collisions of self-driving vehicles can often be reduced to combined failures in the hardware and software components of their underlying systems. Incidents like this seriously challenge the design of reliable and secure systems that operate in the real world and can cope with unpredictable external phenomena and error-prone technology. How should one look at the issues at stake here, from a design philosophical viewpoint?

* Version dated: July 8, 2020. The research of the first author was partially supported by ICS AS CR fund RVO 67985807, programme Strategy AV21 "Hopes and Risks of the Digital Age", and the Karel Čapek Center for Values in Science and Technology. This paper is the full version of [23].

The systems are examples of *cyber-physical systems*, the modern breed of integrated information and communication systems that almost exclusively rely on programs and computers (processors) to govern the behavior of a designated set of processes or mechanisms in the physical world. Cyber-physical systems comprise both the computers and the parts of the physical world they govern [2, 9]. In a more general conception, human operators may be included as components as well [20].

Example 1. Examples of cyber-physical systems include ATM's, heart pacemakers, mobile phones, smart TVs, driverless cars, aircraft, trains, lifts, cranes, power plants, ships, orbital space stations, manufacturing systems, and many other systems. (Cf. [9].) Cyber-physical systems are usually interconnected with many other computerized systems and services.

The development of cyber-physical systems is rapidly progressing with the use of advanced AI techniques. In fact, cyber-physical systems can be seen as generalized robots that are controlling complex, and even critical tasks with as little human intervention as possible. Their design must satisfy the highest standards in reliability and safety. Yet, incidents like mentioned above continue to occur. A natural question is how this state of affairs could be improved.

From a philosophical perspective, the vulnerability of cyber-physical systems is rooted in their lacking or limited cognitive abilities. Indeed, one may well argue that today's cyber-physical systems still operate as 'zombies' when it comes to adjusting to new or varying environments, their 'awareness' of operation and cooperation failures (of their components or as a whole), and reacting properly to combined malfunctions of their modules.

This suggests that cyber-physical systems should be designed with more advanced options of self-control and intelligence, to cope with their complex environment. Even if we do not know how to endow these systems with the facilities of full-blown intelligence or even consciousness, and perhaps one might not even want to (cf. [4, 17]), one can imagine to equip them with important aspects of awareness and behavioral knowledge of the parts of the world that they perceive via their sensors. As Shanahan [5] notes: *We can imagine building something that has some aspects of consciousness and lacks others.*

In this paper we argue that the design of adequately self-controlled cyber-physical systems must guarantee, what we call, *minimal machine consciousness* (MMC), a concept expressing that the systems must operate and act based on, and maintaining, full awareness of (the state of) their components and the situation in their environment. We also propose a *design philosophy* for 'cognitive' cyber-physical systems that fits this goal, based on analogies with the design of weak AI's.

The concept of minimal machine consciousness has emerged in our recent effort to give a practical model for conscious behaviour in robots, based on the theory of automata [22]. The concept was initially meant to provide an exploratory, theoretical approach to the computational modeling of certain basic aspects of (artificial) consciousness that are of interest from a theoretical and philosophical point of view. However, we will show that the underlying ideas can *also* be used in the design of reliable and secure cyber-physical devices that operate in the real world.

We contend that designing cyber-physical systems to be minimal machine conscious is the key to obtaining trustworthy systems, as it strengthens their behavioral flexibility in new or varying environments and their resilience to operation or cooperation failures of their components or as a whole. As a design objective, the notion of

minimal machine consciousness seems to provide the missing link to obtaining safe cyber-physical systems, and it should therefore be applied wherever possible and appropriate. This is summarized in the following *design philosophy* [23]:

> *All cyber-physical systems operating in a given environment, with or without human aid, must be designed as minimal machine conscious cognitive systems.*

Outline In the remainder we describe the essence of the design philosophy we propose. In Section 2 we outline the architectural basis of the cyber-physical systems that will be termed 'cognitive'. (We restrict our treatment here to finite-state systems.) We present the *Sense-Analyze-Compute-Act* paradigm for modeling their operational cycle, in analogy to the feedback loop concept of *autonomous systems* [12].

Then, in Section 3, we define the *four principles* of minimal machine consciousness. We argue in detail why cognitive cyber-physical systems have what it takes to satisfy the criteria, relating MMC to the four stages of their operational cycle. We explain how additional features like *artificial emotions* may articulate aspects of their behaviour. The model derives from the general framework in [22].

In the subsequent sections we argue why minimal machine consciousness gives us a proper tool for the design challenge we posed. In Section 4 we discuss what is typically made possible by minimal machine consciousness and what it amounts to as a design objective. Next, in Section 5, we describe a test for determining whether a cognitive cyber-physical system is indeed minimal machine conscious. We also consider the generalized concept of 'minimal *collective* machine consciousness' for cyber-physical meta-systems such as *cyber-physical human systems* [20].

Finally, in Section 6 we discuss the viability of realizing cognitive cyber-physical systems by means of current software technology and assess the potential of minimal machine consciousness for becoming one of the defining system attributes of *Industry 4.0*. In Section 7, we offer some conclusions. This paper is the full version of [23].

2 Cognitive Cyber-Physical Systems

Our aim is to give a design philosophy for cyber-physical systems that are fully and adequately *self-controlled*, i.e. systems (or their software) that are 'aware' of the internal and external factors that influence and determine their operations and the effects they cause. The philosophy should lead to an easy way to impose all necessary requirements of reliability and safety.

The architectural model of cyber-physical systems as we envision them is that of *cognitive* systems. In this section we describe the model, by sketching its components and operational paradigm. In the next section we argue that, in order to be self-controlled, cognitive cyber-physical systems must be minimal machine conscious.

2.1 Architecture

In general, a cyber-physical system is an embedded entity of (hardware and software) components that is producing a behavior in some environment, based solely on the inputs from its sensory and motor modules and other designated interfaces to the outside world. Some systems may also take inputs from human operators. See [1] for a general introduction to the foundations of cyber-physical systems.

In *cognitive* cyber-physical systems, a specific structure as well as a fitting operational paradigm are imposed that allow for a qualitative assessment of the information obtained from all input sources, and a determination of the appropriate actions that the system should take in response to it.

The basic architecture of a cognitive cyber-physical system C consists of five main parts: its network interfaces, its sensory units, its motor units (or effectors), its finite-state control unit and its dashboard. For simplicity, we will consider all external inputs (e.g. from network interfaces) as inputs from sensory units. See Figure 1 for a typical systems view.

Sensory units Sensory units are devices, modules, or subsystems whose purpose is to detect and register events or changes in the system's environment and send the information about them to the control unit of the system. A sensory unit (sensor or interface) sends both a *representation* of the occurrence of a phenomenon or network event it is specialized to and for sensors, depending on the type of sensor, also a *feedback signal* representing the *accuracy* of the corresponding sensation. The accuracy of a sensation can be graded according to some scale (such as insufficient, low, fair, excellent, and so on) and depends on the nature of that sensation. For example, it can be its magnitude, intensity, frequency, blurriness, etcetera.

Motor units Motor units are devices, modules, or subsystems whose purpose is to perform one or more actions in the environment, seen as components of the system's behavior. Some motor units may serve for the positioning of sensors or of the system as a whole, others may be designed for manipulating various effectors of the system. Motor modules send feedback to the control unit in the form of *reports* stating whether, or to what extent, an intended operation could be realized. The feedback 'accuracy' and the 'reports' together are called the *quality* of the respective feedback.

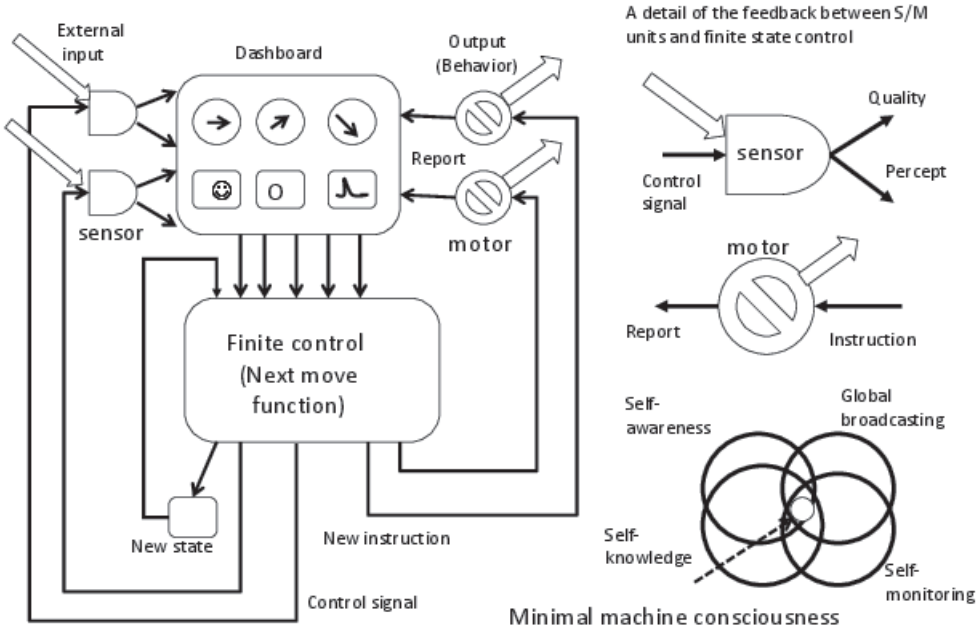

Fig. 1. A schema of a cognitive cyber-physical system

The qualities of the sensations and the reports from the motor modules provide important feed-back information to a system. The graded responses allow the system to monitor the working of its sensory and motor modules. Clearly, not all effectors must perform mechanical movement. Some of them may be 'transmitters' that just produce internal or external signals of some kind: optical, chemical, acoustic, tactile, visual, radio-magnetic, etc. The emitted signals are used for communication purposes, under the assumption that the system and its environment possesses receivers for the respective signals

Finite-state control The finite-state control unit is the *computational* heart of any cognitive cyber-physical system. It acts in a similar way as a deterministic finite-state automaton [22]. The control unit iterates the *operational cycle* of the system (see below). In any iteration, its purpose is to determine the (next) set of actions of the entire system based on four ingredients: the *current state* of the control unit, the current sensations (inputs) from the sensory modules, the quality of these sensations, and the current reports from the motor modules. Typically, the finite-state control unit will be a *multiprocessor* that is programmed to generate the instructions for the set of actions to follow. States represent the possible *configurations* of the unit.

Presuming that there is but a finite number of sensory and motor modules and that the control unit can recognize but a finite number of signals of various types received from these modules, a control unit can produce but a finite number of different instructions. Each such instruction states for a specific sensory or motor module what it has to do 'in the next step'. The instructions may require repeating a previous action, or performing a new specific action, or doing nothing at all. The number of different actions can be very large, even exponential in the number of received signals. (In general, control units may well be allowed to operate on potentially infinite state-sets, but we do not consider this here.)

Dashboard The four ingredients on which the control unit operates are jointly called the *dashboard* information of C. (At any time t, this information can be seen as the *instantaneous description* of the system at time t.) We note that all sufficiently complex (cognitive) cyber-physical systems have some form of 'physical' dashboard that has no influence on computation but is only used by a human operator for keeping a system's behavior within reasonable 'boundaries'. This human activity can be partially or fully automated, as we expect it is in self-controlled systems.

Note that, in principle, the control unit works orders of magnitude faster than many of the other modules, especially the mechanical ones, of a cognitive system. Therefore the entire system works in an asynchronous manner, with its components linked rather as in a distributed system.

2.2 Operation

The second characteristic feature of a cognitive cyber-physical system is its particular *operational cycle*. It is a variant of the well-known 'robotic paradigm', i.e. the *Sense-Think-Act* or *Sense-Plan-Act* cycle, but now with four phases that are iterated in sequence: *Sense-Analyze-Compute-Act* (SACA). The 'Analyze-Compute' part may be seen as a refinement of the 'Think' or 'Plan' phase in the standard robotic case. The four phases are characterized as follows.

Sense In the first phase ('sense') the dashboard information is retrieved, in parallel, by the control unit. The dashboard information must be read in parallel since the next 'step' of the system must be based on all available information at the time a new iteration of the operational cycle begins.

Analyze In the second step ('analyze'), the dashboard information and its gradings are interpreted and fitted against the state information of the finite control, so as to determine how the system and its actions are progressing internally and, naturally, in the system's environment (as far as it can tell from its sensory input). The phase leads to a possible amendment of the current strategy of the system, including the modification of relevant analysis parameters.

Compute In the third step ('compute'), the system's *transition function* is applied to the dashboard information and the current or amended strategy of the system. This is a function that for any given current state, current dashboard information, and the current analysis of it determines a new state of the control unit and for each sensory and motor module, the instructions needed to trigger and realize (or modify) an appropriate action in this iteration.

Act Finally, in the fourth step ('act'), the new state of the control unit and the instructions for the new actions are broadcast to the respective modules in parallel, based on the result of the transition function.

After the modules perform their new operations, a new bundle of data is gathered to refresh the dashboard: new sensations and their qualities from the sensory modules, and new reports from the motor modules. Then the entire operational cycle is repeated. The schema of a cognitive cyber-physical system is depicted in Fig. 1. Speedy operation of a system may require that consecutive iterations overlap in practice.

The *Sense-Analyze-Compute-Act* cycle concept resembles that of the *Monitor-Analyze-Plan-Execute-(over-shared-)Knowledge* (MAPE-K) feedback loop as known in the design of self-adaptive *autonomic systems* [12]. A formal description of the entire model can be given in a framework like suggested in [1] or in the same way as in the automata-theoretic framework described in [22].

Definition 1. *A cognitive cyber-physical system is called* complete *if and only if its transition function is defined for all combinations of its inputs.*

A complete cognitive cyber-physical system can, in principle, react differently to different inputs in response to changes in its input parameters.

3 Self-controlled Cognitive Cyber-physical Systems and Minimal Machine Consciousness

As cognitive cyber-physical systems operate, potentially, in full 'cognizance' of their components and their environment, it makes sense to consider how they might operate under full and adequate self-control. In this section we first outline how the four-phased operational paradigm leads to the four principles of self-control which we collectively call *minimal machine consciousness* (in analogy to [22]). Next we argue that all self-controlled cognitive cyber-physical systems must be minimal machine conscious. We conclude with several further reflections, e.g. on articulating aware behaviour through 'artificial emotions'.

3.1 Minimal Machine Consciousness

Cognitive cyber-physical systems can be adequately self-controlled only when a full 'picture' of itself and its embedding can be derived (implicitly), based on the information from their sensory and motor units and on the potential actions they can initiate at a given moment.

Considering the *Sense-Analyze-Compute-Act* paradigm, it makes sense to distinguish four corresponding 'dimensions' that together serve as prerequisites for adequate self-control. This leads to the following 'self-$\star$' properties which one might consider for a cognitive cyber-physical system C:

- *self-knowledge*: C has complete knowledge of its current cognitive state as well as of the data produced by all its interfaces, sensors and motor units.
- *self-monitoring*: C is completely informed about the performance and status of its sensory and motor units over time (including the quality of the sensations and the reports from all of them) and of its embedding in the environment as it is.
- *self-awareness* (or *self-reflection*): C behaves in a way that unambiguously reflects, resp. is determined by, its current cognitive state and the information gained by its self-knowledge and self-monitoring abilities, and that is 'aware' of the internal and external changes that it causes.
- *self-informing*: C globally broadcasts its cognitive state, to all modules of the system and whenever changes of state occur.

Note that these self-$\star$ properties are all oriented towards the 'cognitive' behaviour of a cyber-physical system, as opposed to the self-$\star$ properties normally distinguished for autonomous systems (self-configuring, self-optimizing, self-healing, self-protecting) which are oriented towards the self-management of a system's software [12].

Definition 2. *A cognitive cyber-physical system C is called* minimal machine conscious *(MMC) if and only if it is self-monitoring, self-knowledgeable, self-aware, and self-informing.*

There are several reasons for using the term 'minimal machine consciousness' for the collective properties we distinguished. A major reason is that, together, they seem to represent the minimum requirements for a system to be *non-zombie*: omitting any of these requirements might cause a system not to respond adequately under all circumstances it may encounter (cf. Section 5). Furthermore, a cognitive cyber-physical system was defined as a finite-state system, without relying on further resources. This means that the 'active' memory available to realize any sort of 'conscious behaviour' is only assumed to be finite, i.e., 'minimal' when compared to intelligent systems with (potentially) unbounded active memory.

The four principles of self-control are necessarily *informal*. We envision that for any class of cyber-physical systems they are concretized, to the extent that they provide precise requirements for the system designers and are verified for the systems that are claimed to satisfy them.

Example 2. Several disasters of airplanes and space shuttles have been caused by the lack of self-knowledge and self-monitoring qualities, and the absence of cooperation among the modules of the flight-control system. For example, in 1986, the space shuttle Challenger exploded due to an unspotted malfunction of the spacecraft's rubber seals. No one on board survived. The recent crash of Lion Air Flight 610 was caused by a malfunctioning of the flight-control system of a Boeing 737 MAX 8 that should not have happened if it had been a minimal machine conscious system.

Example 3. Minimal machine consciousness is not necessarily restricted to cognitive cyber-physical system that have a finite number of configurations only. For example, note that a Turing machine can be seen as a cognitive cyber-physical system in which a finite-state control governs a finite set of sensory and motor units, namely the respective read/write heads on its worktapes. Operating over an input stream by means of the actions and feedbacks from the read/write heads, the system is seen to be minimal machine conscious. Nevertheless, the work-tapes give it a potentially unbounded memory and thus, an unbounded number of configurations. In general, if a system has potentially unbounded memory at its disposal, it can store information about all events it ever experienced in the past and is currently experiencing, giving it far greater 'cognitive' abilities, at least potentially, than the cyber-physical system we consider.

3.2 Self-controlled Cognitive Cyber-physical Systems are MMC

The four principles that define minimal machine consciousness (self-knowledge, self-monitoring, self-awareness and self-informing) correspond precisely to the properties that are required for full and adequate self-control in the various phases of the operational cycle of a cyber-physical system. We formulate this as follows.

Proposition 1. *Self-controlled cognitive cyber-physical systems are necessarily minimal machine conscious.*

We expand on our arguments in support of the proposition below, considering each of the properties of minimal machine consciousness in turn:

(a) Self-controlled cognitive cyber-physical systems have self-knowledge

The information needed for self-knowledge includes its current cognitive state and the information produced by its interfaces, its sensory units and its motor modules. In a cognitive cyber-physical system, this is part of the data maintained in the dashboard. It gives the system the possibility to report any information that is needed about its functioning, at any time.

(b) Self-controlled cognitive cyber-physical systems are self-monitoring

In the state it is in, the feedback from the sensory and motor modules as it is supplied by the feature of self-knowledge, makes it possible for the system to monitor itself. Namely, from these modules the system gets the data about its current working conditions (the qualities of the sensations and reports of the motor units), and based on this information it can either prolong its functioning without any further special actions or take steps that remedy or adjust its operation. All this confirms the machine's certainty, or errors, in its actions and enables the repair of its own mistakes [3].

(c) Self-controlled cognitive cyber-physical systems are self-aware

The current cognitive state and the information gained by its self-knowledge and self-monitoring abilities, enable a system to determine ('compute') whatever its appropriate next action would be, in the environment in which it operates. The property of self-awareness follows from the fulfillment of the following three conditions:

– *the capacity of introspection*, i.e. the ability to reflect on one's own mental state (cf. [3]). General mechanisms of introspection seem to be beyond the ability of finite-state devices, as they may require unbounded memory. In the framework of finite-state systems, introspection can be modeled by a finite number of system states. For instance, 'interesting' past states can be stored ('remembered') in the current state, by using the standard automata theory technique of storing data in an automaton's state. In this way, one can even introduce dedicated states of the control unit, so-called

machine qualia states, in which a system can remember important past events that still require its ongoing attention (cf. [22]). Machine qualia offer the system a mechanism for remembering certain 'subjective' cognitive states of the system that are bound to certain previous cognitive 'experiences' (states). For instance, a quale state in a mobile phone may keep a remembrance of a recent event when a text message was received. A driverless car can have a quale state regarding a shortage of gas. The qualia states stored in the system's global state can then be broadcast to the entire system as long as a circumstance invoking them persists.

– *the ability to recognize oneself as an individual object separate from the environment and other objects*. This will be implied by a proper selection of sender-receiver modules whose cooperation provides the required effect. There are several modalities of signals that can have a similar effect. For instance, receiving a specific olfactory (or chemosensory), electric, optical, acoustic or haptic return signal may indicate the presence of other instances of the system. Obviously, the absence of such return signals indicates that no similar systems are around. For a similar purpose, in advanced cyber-physical systems a vision system may be available.

– *awareness of changes in the outside world*. The feedback also allows the system to distinguish its actions as registered by its sensory modules from the similarly registered actions performed by other systems. That is to say, in the latter case, the reports from the motor modules do not match the sensations from the sensory modules.

Self-awareness thus provides a cognitive system with a rudimentary machine concept of the *self*: the system has information on what goes on in the outside world, what its actions are and what their effects. This information is of the form 'here and now' – it is pertinent to the present position of the system in its environment and the present moment.

(d) Self-controlled cognitive cyber-physical systems are self-informing

By the very definition of cognitive cyber-physical systems, the new current state of a system and the projected actions are 'broadcast' to all its modules, simultaneously and in parallel. This ensures a synchronization of the actions when needed and gives the modules a certain minimal information (namely, that 'stored' in the current state) of what goes on in the entire system. Endowing machines with the possibility of self-informing allows their modules to share information and collaborate, to address whatever impending problem (cf. [3]). For example, consider a modern car in which a fuel sensor reports a shortage of gas. If this information is globally available, then the navigation system of the car can direct the driver to a nearest gas station [3].

Proposition 1 shows that minimal machine consciousness is a necessary condition for adequate self-control of a cyber-physical system. This gives a technical guideline for the desired behaviour of a system.

Example 4. If a given (cognitive) cyber-physical system does not satisfy all necessary criteria of MMC, it is instructive when analyzing a system's behaviour to find out what properties are missing that would make it minimal machine conscious. For instance, driverless cars still seem to lack the full property of 'self-informing'. (For example, currently, their GPS unit is not informed about an imminent malfunctioning of the engine, which means that the car may not be directed in time to the nearest service station.) Similarly, deficiencies concerning their insufficient orientation in traffic point to insufficient 'self-awareness'. This type of analysis can point to shortcomings in a system's design even before testing it.

In Section 4 we will see that minimal machine consciousness leads to a variety of further properties that are normally associated with self-control.

3.3 Reflecting on the Notion of Minimal Machine Consciousness

A cyber-physical system will not only be programmed to achieve a certain task, but also to achieve it optimally under the varying circumstances that it may encounter. One may well ask whether, and how, minimal machine consciousness may be utilized to deliver this type of extra control as well.

If a cognitive cyber-physical system is minimal machine conscious, thus self-monitoring and self-aware, it is potentially able to respond to any (internal or external) event that comes on its way. However, if this event threatens to interfere with the ongoing task in an critical way (whatever this means), the system may be 'motivated' to call on special routines to salvage or restore the system's mission. This results in so-called *artificial* (or: *machine*) *emotions*, which cause the system to evaluate its options and respond by modifying its behaviour or performance parameters.

Machine emotions are reactions of the physical components or changes in the behavior of a cyber-physical system to special classes of internal and/or external signals. The notion originates from the seminal work of Simon [18] who describes (human) emotions as 'interrupt mechanisms' that enable a (cognitive) system to react quickly to situations that require urgent response. It may take a while before the system can return to normal mode. We refer to [7] for a further exposition of the theory of emotions as rational responses of a cognitive system.

It thus appears that, if cognitive cyber-physical systems are designed to be minimal machine conscious, the notion of machine emotions can be utilized as a programming abstraction 'free of charge' in its design as well. The idea and use of emotions in artefactual systems is a well-studied topic for many years [19].

4 A Manifesto On Minimal Machine Consciousness

We claim that minimal machine consciousness is a key criterion for all cognitive cyber-physical systems in practice, in order to provide these systems with the necessary abilities for smooth and safe operation. It has important consequences for the engineering of cyber-physical systems. We summarize this in the following assertion.

> *All cyber-physical systems operating in a given environment, with or without human aid, must be designed as minimal machine conscious cognitive systems.*

In this section we argue for the importance of minimal machine consciousness. We also *hypothesize* that cognitive cyber-physical systems can always be *re-designed* so as to be minimal machine conscious.

4.1 What Minimal Machine Consciousness Makes Possible

In Table 1 we list a variety of important abilities which cyber-physical systems should have and the mechanisms that facilitate them if the system is (cognitive and) minimal machine conscious. It shows what is made possible by the combined properties of the architecture and the four principles of minimal machine consciousness.

The feedback from the sensory and motor units brings straightforward benefits for improving system performance. Self-knowledge, self-monitoring, and self-awareness lead to improved decision making and increased detection capabilities. Self-informing

enables the cooperation and synchronization of a system's modules. Altogether, minimal machine consciousness enables a system to detect and correct failures that can potentially prevent a possible crash or disaster, and at least diminish the number of false alarms, thus improving the trustability, reliability and safeness of the system under the changing conditions in its environment.

Ability	**Mechanism**
Improved decision making Increased detection capabilities Diminished number of false alarms Failure correction Damage registration Flexibility and improved reliability in varying situations	Graded feedback from sensors and motor units
Interception of adversarial physical actions	Additional sensors
Attention mechanism	Suppressing disturbing inputs
Limited form of introspection	Cognitive states
Detection of patterns in ongoing processes	Introspection
Recognition of itself as an individual subject, separate from the environment and other systems	Cooperation of send-receive mechanisms
Communication	Send-receive mechanisms
Distinguishing one's own actions from the actions of other systems	Mismatch of motor actions with sensory observations
Reading the intentions of other similar systems (machine empathy)	Situating the system into the position of the other systems
Limited cognitive and calculatory tasks	Finite-state data processing
Subjective machine perception (machine qualia states)	Storing states in states

Table 1. Abilities of minimal machine conscious cyber-physical systems and the corresponding mechanisms that realize or facilitate them

The detailed mechanisms of self-awareness lead to further potential benefits. For instance, self-awareness requires that the system must be able to distinguish its own movement from any other movements that it can observe in the environment. This property can be used, e.g., by a robotic arm system to intercept a motion (of an 'intruder') within reach of its arm. Another example is collision-free navigation. As an extreme case, a minimal machine conscious system can 'read the mind' of another, similar system by observing its input and by being aware that this is not its input. Namely, thanks to the fact that the observing and the observed systems are of the same construction, the observing system can infer the actions of the observed system.

One may note also that minimal machine consciousness is a prerequisite for *autonomous behavior*, and thus for *ethical behavior*. Obviously, if a cognitive cyber-physical system would not fully adhere to the four principles of minimal machine consciousness, it would not always react adequately to all sensory stimuli, or make use of introspection, or recognize itself as an individual subject, separate from the environment and other systems, or distinguish its own actions and their consequences from the actions of other agents. Hence, it would not behave rationally, and the more so, autonomously and ethically in every conceivable situation.

4.2 Design Considerations

As presented, minimal machine consciousness becomes feasible once a cyber-physical system is, or can be, designed as a *cognitive* system. This follows from the close con-

nection between the four principles of minimal machine consciousness and the necessary features of *self-control* during the consecutive phases of the operational cycle of the system. Minimal machine consciousness enables the system to operate awarely in its environment at any time.

We therefore contend that minimal machine consciousness should be one of the major *design objectives* of any cyber-physical system. Having this is mind, the following bold statement is at the heart of our design philosophy.

Claim 1 *Any cognitive cyber-physical system operating in a given environment, with or without human aid, can also be designed as a minimal machine conscious cyber-physical system.*

To see this, consider any cognitive cyber-physical system, operating in a given environment. By design, the system will have the knowledge of what behavior must be invoked, based on the continued inputs from its modules to the dashboard. It should thus be possible to utilize this information from the dashboard fully and 'redesign' (or, re-program) the system, so as to transform it into one that conforms to the four principles of minimal machine machine consciousness described above.

We go even a step further when it comes to the *architecture* of cyber-physical systems. This concerns the modeling, construction and programming of any cyber-physical system that mimics the activity in some domain of conscious human behavior in general. We contend that these systems can always be designed (or, re-designed) so as to conform to the architecture and operation of cognitive cyber-physical systems. We therefore *hypothesize* the following, stronger claim.

Claim 2 *All dedicated activities that can be consciously controlled by humans can also be controlled by minimal machine conscious cognitive cyber-physical systems.*

The background for this claim is that, if we take 'conscious control' to mean as much as 'possessing knowledge of how to behave to fulfill a certain task', then one must be close to knowing or discovering the dependencies between the various components of the required behavior and the corresponding inputs from the interfaces and the sensory and motor modules. If we accept the cognitive architecture as a *standard*, then the rules and necessary information to drive the operational cycle should be in reach.

The claims can be the starting point for further methodological, or software engineering considerations towards the realization of cyber-physical system as described in, for example, [9] and [10]. Including minimal machine consciousness as a concrete design objective calls for more orderliness and discipline in the design of the system, by insisting on the fulfillment of the four necessary conditions required for this type of 'consciousness'. The benefits are clear. The starting points for the methodology will be further refined in Section 5.

Example 5. It seems that, currently, no clear-cut example of a deliberately designed minimal machine conscious cyber-physical system exists. The closest example seems to be the modern smartphone. These phones usually have a 'dashboard' in the form of a *status bar* (e.g. along the top of the touchscreen). Here, the statuses of various important sensors, such as the quality of the wi-fi, mobile network, GPS, the bluetooth signal, battery power, etcetera, are depicted with the help of the respective icons. At the same time, the icons indicate the quality of the respective signals. This is, in fact, global information describing the current state of the device that is accessible to all modules of the phone, and a witness of the system's self-knowledge and self-monitoring. Last but not least, the system is quite self-aware

– it can recognize the incoming calls, send and receive messages, establish the bluetooth connection, identify changes in its location (via GPS), etcetera. Interestingly, these abilities of smartphones are the result of an incremental, technological evolution and the development of user requirements rather than of a purposeful effort to make the devices minimal machine conscious. This only confirms that the idea of minimal machine consciousness is a natural and useful concept that is worth to follow up and exploit as a design objective.

In the sections below we will give further arguments in favor of minimal machine consciousness as a required feature of adequately designed (cognitive) cyber-physical systems.

5 Qualities of Minimal Machine Conscious Cyber-Physical Systems

In this section we consider some further qualities of (cognitive) cyber-physical systems. First we discuss the importance of feedback for achieving MMC. Then we elucidate what distinguishes minimal machine conscious cyber-physical systems from many currently known systems, including so-called *zombie systems*. Based on this, we propose a *test* for minimal machine consciousness. Finally, we argue that minimal machine conscious cyber-physical systems are perfectly geared to 'cooperate' with other systems of the same caliber, leading to a refinement of our design philosophy for cyber-physical systems. The methodology even applies when humans are included as components, in so-called *cyber-physical human systems* [20].

5.1 Feedback and MMC

A cognitive cyber-physical system operates with two kinds of inputs (signals). The first kind results from its self-knowledge and consists of *direct sensations* from its (interfaces and) sensory modules, in some represented form. The second kind results from its self-monitoring ability and consists of *feedbacks* both from the sensory (their qualities) and the motor modules (the reports). Based on the direct inputs and the feedback information, the control unit adjusts the cognitive state and the system's behavior.

Note that cyber-physical systems that would operate solely based on the direct inputs would, in fact, obtain insufficient data to drive the components of the systems safely and securely. It could jeopardize the system's behavior since the inputs may be incomplete or flawed, for various technical as well as environmental reasons. However, properly designed software that makes use of the feedback information will enable the cognitive system to take appropriate actions.

It is exactly this difference between the behavior with and without feedback information, that captures the distinction between a system that is or is not minimal machine conscious. Feedback information is necessary for a system's *autonomous behavior*, as noted earlier. After all, how could a system be autonomous and always react properly, if it does not register all its internal or external stimuli all the time?

5.2 Testing for Minimal Machine Consciousness

An interesting question is how one might be able to tell whether a given (cognitive) cyber-physical system is minimal machine conscious. To this end we first consider those systems that lack one or more of the self-control options of minimal machine consciousness. Among current cyber-physical systems, they are the common type.

Definition 3. *A cognitive cyber-physical system C is called a* zombie system *if and only if C is not minimal machine conscious.*

Zombie systems are either not fully self-knowledgeable, self-monitoring, self-aware or, indeed, not fully self-informing. If any of the four abilities of MMC is not fully present, then situations may occur of which the characteristics are not completely (or not correctly) registered by the respective mechanisms of the cognitive system (viz., on its dashboard). Consequently, the system may not behave or react properly in these situations, possibly causing damage or disaster. This is the basic idea behind the following assertion, expressing that minimal machine conscious cyber-physical systems have superior self-control compared to zombie systems (cf. [22]).

Proposition 2. *Let $\mathcal{M}$ be an arbitrary minimal machine conscious (i.e., non-zombie) cyber-physical system and $\mathcal{Z}$ an arbitrary zombie system. Then there exist situations in which $\mathcal{Z}$ cannot behave in the same way as $\mathcal{M}$ does.*

Clearly, under the assumption that all sensory and motor modules of a cognitive cyber-physical system work flawlessly and the environment does not put any obstacles in the way of its normal operation, the system does not need the facilities of minimal machine consciousness. In this case its tasks may as well be performed by a zombie system. Minimal machine consciousness is required as soon as feedback information from the modules becomes essential for the system's flawless operation.

In essence, Proposition 2 expresses that *the class of cognitive tasks that can be handled by minimal machine conscious systems is strictly larger than the class of tasks handled by zombie systems*. Understanding the behavior of zombie systems is important for determining how one might actually *test* whether a given cognitive system is minimal machine conscious.

Testing Based on Proposition 2 and its justification, an effective test for minimal machine consciousness can be given which can be applied to all cyber-physical systems that sensibly qualify and allow for experimentation (e.g. using a *simulator*). This is important both for existing systems and for new systems at relevant stages of their design (presumably, as a minimally machine conscious system).

Proposition 3. *Let C be an arbitrary cognitive cyber-physical system that is constructed so as to behave purposefully under all internal and external conditions that it can encounter. Then one can effectively test C for minimal machine consciousness.*

Given a cognitive system C, the idea of the test is based on the common practice in *software and system testing*, namely to check experimentally whether the assumptions on C's operation are always fulfilled. To this end, one tests whether C indeed adjusts its behavior as desired to all variable conditions in its environment that could complicate its mission. That is, one checks whether C continues to operate meaningfully whenever the system, its interfaces, and its sensors or motor units are (temporarily) interfered with. If, under any such conditions, C starts to behave erratically or nonsensically, we conclude that it is not minimal machine conscious and thus a zombie system.

If C is a zombie system, then it cannot pass the test, if the testing is sufficiently exhaustive. After all, by Proposition 2, there will be situations in which C misses some information that is important for adequate behavior due to the lack, or insufficient mastering, of the full effect of minimal machine consciousness.

When testing for minimal machine consciousness, one first has to know whether the hardware is fully reliable, i.e. whether its interfaces, sensors and effectors work as required, whether they react properly to the signals sent by the finite-state control unit, whether they return properly graded feedback, and whether their constellation is fault-tolerant. Only then does it make sense to test that the system always adheres to its operational goal. That is, whether its *behavior*, i.e. the sequence of actions as produced in any given environment and under any given circumstance, is 'meaningful' and 'non-erratic', i.e. corresponds correctly to its mission.

Note that the experimental test can be used in real situations, e.g. for testing whether a flight control system can cope with all situations when critical parts of the system may be malfunctioning. Even if a system has not been designed as being minimal machine conscious, the criteria of MMC are important for the design of properly focused test procedures.

Example 6. Consider again the disaster of Lion Air Flight 610 in October 2018. According to Wikipedia, already in earlier flights, and during the flight itself, several malfunctions of the aircraft's control system were reported. Due to insufficient knowledge of the situation and its causes, and lacking the training for such a situation, the pilots were not able to react properly to intercept the looming causality. In our terminology, the cyber-physical system represented by the aircraft, its on-board computers, machinery, and the crew, was not minimal machine conscious. All information was available, so-to-speak literally on the dashboard, but what was missing was the full power of self-informing: making the globally available information about the state of the affairs known to all parts of the system. Even if this information would have been available, the control system, inclusively the pilots, was not 'programmed' to deal with the situation. This is because the system was not tested for completeness. A properly designed system with minimal machine consciousness might have resolved the situation even without human aid.

Testing the four criteria of minimal machine consciousness for cognitive cyber-physical systems is not merely a matter of software and system testing, but notably of *requirements analysis* and *completeness testing*. Their dual role in requirements engineering is well-known.

Finally, all further insights from the theory of testing apply. In particular, no amount of testing normally suffices to guarantee that the criteria of minimal machine consciousness are always satisfied: it is an empirical fact that there will always be circumstances in practice that are not anticipated in a testing process. *This does not contradict our manifesto but, rather, reinforces it as a design philosophy.* It does support the principle that, ideally, all critical parts of a system are *formally verified*, based on the detailed specification of its MMC properties.

On consciousness testing Testing cognitive systems for minimal machine consciousness reminds of the well-known problem from the Theory of Mind whether there is a test for *consciousness*. In philosophy, possessing (human-like) consciousness is often seen as defining an important boundary between behavior that is 'responsible' and behaviour that is not, for animals *and* robots (cf. [17]).

The principle of our test for minimal machine consciousness is equally suitable for consciousness testing, for living as well as non-living entities. Compared to the test recently proposed by Schneider and Turner [16, 17] for detecting consciousness in robots (a variant of the Turing test that focuses on discovering subjective experiences during a verbal interaction with the tested subject), one immediately sees the advantages of our approach. Namely, it is a behavioral test not dependent on understanding a natural language.

5.3 Minimal Collective Machine Consciousness

Given a number of different cyber-physical systems, there may be considerable potential in combining them into one 'composed system'. This happens, for example, when a complex task must be split over several cyber-physical systems, with each system dedicated to a well-identified subtask. This leads us to consider *networks* of cooperating cyber-physical systems that all handle different subtasks and collectively perform the task at hand.

If the 'nodes' are all cognitive cyber-physical systems, then one may turn the network into a cognitive cyber-physical 'meta-system' by adding a global finite-control unit that sees the nodes as sensory/motor units and combines their information into one global operational cycle (which need not be synchronized with the operational cycles of the nodes). This construction is especially interesting when the nodes are all minimal machine conscious. The network will, in general, be *self-organizing*.

Example 7. One may think of modern robots as cognitive cyber-physical meta-systems, with subsystems dedicated to specific tasks like vision, motion, sensing, and grasping. More generally, teams of robots, swarms of drones, nano-machines in a bloodstream, etc., also qualify.

To see what sort of additional machine consciousness this may lead to, consider an arbitrary cyber-physical meta-system D. The nature of the information D collects from its nodes may vary widely. It can be data from the specific subtasks of the nodes, statistics related to their activities, reports on the working conditions and cognitive states of the underlying cyber-physical systems, etcetera. Based on this, D can keep track of the part of the 'world' that is registered by its nodes. In particular, if the nodes are minimal machine conscious, D may collect their qualia states.

As a result, D has all it needs to register the prevailing 'spirit' of its node systems and is, as we say, *minimal collective machine conscious*: it is self-knowledgeable, self-monitoring, and self-aware of what goes on in the network, and it will be as self-informing to the nodes as is needed and foreseen for their collective operation as a meta-system. Note that, after processing all information, D's finite-control unit need not send the new state and action information to its nodes necessarily in a parallel, 'one-to-all' manner. The appropriate information will normally be broadcast over the connections as they exist in its network at the time.

From an abstract perspective, cyber-physical meta-systems are just cyber-physical systems that are more explicitly tailored to *structured design*, with dedicated cyber-physical systems constituting one integrated system. With the *design philosophy* of cyber-physical system design in mind (cf. Section 4), these considerations now lead us to the following, bold, *refinements* of Claims 1 and 2.

Claim 3 *Any cognitive cyber-physical meta-system with its components operating in a given environment, with or without human aid, can also be designed as a minimal collective machine conscious cyber-physical meta-system.*

Claim 4 *Any complex of dedicated activities that can be consciously controlled by humans can also be controlled by a minimal collective machine conscious cognitive cyber-physical meta-system.*

These claims jointly complement the design methodology for cyber-physical systems advocated in our philosophy.

The notion of (minimal collective machine conscious) cyber-physical meta-system can be extended by explicitly including human agents, in an indirect or active role in the networks. This leads to a powerful bread of systems called *cyber-physical human systems* or *cyber-physical social systems*, which can provide a host of services and tasks (cf. [20, 25]). We quote the following characterization from [20]:

Definition 4. *A cyber-physical human system consists of interconnected systems (computers, cyber-physical devices, and people) 'talking' to each other across space and time and allowing other systems, devices, and data streams to connect and disconnect.*

Cyber-physical human systems may be seen as an interconnected network of cyber-physical systems, together with humans which manage and utilize them. The design methodology of cyber-physical meta-systems directly applies to them as well.

Example 8. As examples, one may consider a smart city, or a small grid, or a (modern) country with e-government, etc. In general, one may think of all kinds of intelligent infrastructures that involve humans and computers (cf. [21], [11]), social networks, and e.g. the social credit system developed in China.

Cyber-physical human systems can be used, but also abused, for many different purposes in society, and their design requires careful ethical control and supervision. When complex cyber-physical systems are knowingly designed as minimal machine conscious systems, it is pertinent that we strive for systems that are not only autonomous, but also provably ethical.

6 What Can We Do?

The design and development of cyber-physical (meta-)systems often involves the most advanced forms of modern AI technology, to create and achieve some desired technological goal. It is tempting to add further capabilities of intelligence and autonomy. Many philosophers before us have spoken out against the development of systems with these capabilities, which may well turn on us in unexpected ways. But then, quoting Dennett [4]: *what can we do?*

In this section we will argue that it is sufficient to strive for minimal machine consciousness as a technological answer.

6.1 Desideratum

A good illustration is the development of autonomous, or driverless, vehicles. The best currently produced driverless cars are classified as so-called *level 1* ('hands-on') cars, where the driver and the car-guidance system share the control of the vehicle (cf. [24]). These cars are already equipped with many functions, like self-parking, speed limiter, cruise control, anti-skid braking, air conditioning, charcoal filter cleaning, customization options, emission test, fuel management, direct download of software updates, active suspension, start-stop, navigation, etcetera (cf. [10]). Cars of level 2 or higher have even more, potentially autonomous functions.

The question is: do all combinations of these functions work properly when they work at the same time? For instance: can start-stop and air conditioning be active simultaneously? Is cruise control compatible with speed limitation? Can self-parking be

activated during a software download? Does the active suspension smooth the impact of a pothole during self-parking? (cf. [10]). Should navigation to the nearest service station be invoked when an engine malfunction is detected? It is unclear whether these forms of self-control are present, and whether the completeness of the system has been sufficiently tested.

In the preceding sections we have argued that the answer is contained in designing car systems as cognitive cyber-physical systems that are minimal machine conscious. A *minimal machine conscious driverless car* should be able to resolve the listed problems without failures of design.

6.2 Minimal Machine Consciousness and Industry 4.0

It is a serious challenge to design a specific minimal machine conscious cyber-physical system that can handle all possible situations, given the numerous situations that the system can face. This is well-recognized in the area of software engineering, for embedded systems and cyber-physical systems alike. However, the challenge must be met, as cyber-physical systems are crucial for all enterprises. This is notably expressed in the advanced concepts of smart manufacturing in *Industry 4.0* [8].

In Industry 4.0 it is foreseen that all production processes in factories are automated and computerized, making them flexible and efficient by the use of modern information and communication technologies and the advanced options of intelligent systems and services [14]. The processes will be connected and controlled by smart systems that manage entire production lines and make decisions of their own, in symbiosis with human operators. We refer to [13] for an overview of the technological challenges involved.

The core systems of Industry 4.0 can be recognized to be cyber-physical (human) systems or even meta-systems as defined in Section 5. As we have argued in this design philosophy, these systems must be designed so as to be 'cognitive' and minimal (collective) machine conscious. With the testability of the latter in mind, this seems certainly achievable when taken into account as a criterion from the outset. It may be the limit of what current hardware and software engineering methods can do.

Fortunately, a promising technology is emerging that enables both the design and the efficient testing of potentially minimal machine conscious cyber-physical systems: the technology of *digital twins.* In our case, a digital twin would be a digital replica of the physical part of a designated cyber-physical system. Using a digital twin of (a part of) the cyber-physical system, one can systematically test whether it will react properly to all external and internal conditions, provided the combinations and scenarios that the system's modules may face can be finitely enumerated. If a system passes such a test, we know that it is complete with respect to all events that can be registered by the system (cf. Definition 1). Of course, if the testing cannot be exhaustive, the system can only be guaranteed as far as the scenarios went.

Digital twin research is a growing and flourishing scientific field (cf. [6]). Its application to the design and testing of minimal machine conscious systems can give further impetus to the research and development of this technology. It is generally accepted that digital twin technology is one of the key enablers of *Industry 4.0*. The concept of minimal machine consciousness has the potential to revolutionize this field further.

The concept brings a new quality to digital technologies, by making them more flexible, more robust, and, last but not least, safer and more secure.

6.3 A New Sort of Entity

Dennett [4] claims that *"We don't need artificial cognitive agents. We need intelligent tools."* His claim seems to contradict the spirit of our design philosophy, that pleads in favor of using minimal conscious artificial cognitive agents.

However, what Dennett means is that we do not need humanoid agents endowed with a human-like consciousness. Rather, what we need are *intelligent tools* that 'do not have rights, feelings, and cannot suffer'. To quote Dennett [4] again, we need *"an entirely new sort of entity, rather like oracles, with no conscience, no fear of death, no distracting love and hates, no personality, but with clear advantages in terms of their user benefits."*

We believe that minimal machine conscious cyber-physical systems follow his vision. Contrary to Dennett [4], we believe that *"it will* not *be hard [. . .] learning to live with [these systems] without distracting ourselves with fantasies about the Singularity in which these AIs will enslave us, literally.*"

7 Conclusion

We analyzed the current complexities of cyber-physical systems and found the concept of minimal machine consciousness to be a possible answer to the problem of making these systems more robust and safe. The main message of this paper is the following assertion.

> *All cyber-physical systems operating in a given environment, with or without human aid, must be designed as minimal machine conscious cognitive systems.*

Minimal machine consciousness is applicable to all systems, not only to highly complex ones. However, it is not an add-on feature that can be achieved by merely upgrading the software. In our approach, it requires a suitable system architecture to deal with the graded feedbacks from all sensory and motor units, in addition to a proper embodiment that suits the system's intended task. Even when these requirements are not met, the criteria of minimal machine consciousness can still be aimed for.

Minimal machine consciousness has a meaningful purpose, its benefits are substantial and can hardly be obtained differently. What is costly, however, is the development of systems that have this property, as maximal attention must be paid to their functionality under all possible conditions, be they caused by software or hardware malfunctioning, or by unfortunate combinations of adversarial external factors. In existing systems, minimal machine consciousness can be a target in the gradual evolution of the software, if the system allows for it.

More attention must be paid to the methodological aspects of developing and testing minimally machine conscious cyber-physical systems, since current methodologies do not take the requirements of minimal machine consciousness into account [9, 10]. When designing new cyber-physical systems, or innovating existing ones, especially those in which risks for human life are at stake, it is the matter of responsible design and ethics to make such systems minimal machine conscious.

References

1. Broy, M.: Engineering Cyber-Physical Systems: Challenges and Foundations. In: M. Aiguier *et al.* (Eds.), *Complex Systems Design & Management*, Proc. Third Int.Conference (CSD&M 2012), Springer-Verlag, 2013, Ch. 1, pp. 1-13
2. Broy, M., Schmidt, A.: Challenges in Engineering Cyber-Physical Systems, *IEEE Computer* 47:2 (2014) 70-72
3. Dehaene, S., Lau, H., Kouider, S.: What is consciousness, and could machines have it? *Science*, 358: 6362 (2017) 486-492
4. Dennett, D.C.: What Can We Do? In: J. Brockman (Ed.), *Possible Minds: 25 Ways of Looking at AI*, Ch. 5, Penguin Press, 2019
5. Edge: The Space of Possible Minds - A Conversation With Murray Shanahan, *Edge*, May 18, 2018
6. ERCIM: Digital Twins: Special Theme. *ERCIM News* 115, October 2018, https://ercim-news.ercim.eu/en115
7. Frijda, N.H.. *The Emotions*, Cambridge University Press, Cambridge UK, 1986
8. i-Scoop: *Industry 4.0: the fourth industrial revolution - guide to Industry 4.0*, March 2020, https://www.i-scoop.eu/industry-4-0/
9. Jackson, M.: Behaviours as Design Components of Cyber-Physical Systems. In: B. Meyer, M. Nordio (Eds.):, *Software Engineering*, International Summer Schools, LASER 2013-2014, Revised Tutorial Lectures, Lecture Notes in Computer Science, Vol. 8987, Springer, 2015, pp. 43-62
10. Jackson, M.: Behaviours and Model Fidelity in Cyber-Physical Systems. In: *Computability in Europe: Computing with Foresight and Industry* (CiE 2019), Special Session: History and Philosophy of Computing, Durham
11. Jordan, M.I.: Artificial Intelligence - The Revolution Hasn't Happened Yet. In: *Medium*, April 19, 2018, https://medium.com/@mijordan3/artificial-intelligence-the-revolution-hasnt-happened-yet-5e1d5812e1e7
12. Kephart, J.O., Chess, D.M.: The Vision of Autonomic Computing, *IEEE Computer* 36:1 (2003) 41-50
13. Lu, Y.: Industry 4.0: A survey on technologies, applications and open research issues, *Journal of Industrial Information Integration* 6 (2017) 1-10
14. Rüßmann, M., Lorenz, M., Gerbert, P., Waldner, M., Justus, J., Engel, P., Harnisch, M.: *Industry 4.0 - The Future of Productivity and Growth in Manufacturing Industries*, Report, Boston Consulting Group, April 2015
15. Schneider, S.: It May Not Feel Like Anything To Be an Alien: Humans may have one thing that advanced aliens don't: consciousness. *Nautilus*, December 2016, http://cosmos.nautil.us/feature/72/it-may-not-feel-like-anything-to-be-an-alien
16. Schneider S., Turner E.: Is Anyone Home? A Way to Find Out if AI Has Become Self-aware. *Scientific American*, Observations, July 19, 2017, https://blogs.scientificamerican.com/observations/is-anyone-home-a-way-to-find-out-if-ai-has-become-self-aware/
17. Schneider, S.: *Artificial You: AI and the Future of Your Mind*. Princeton University Press, Princeton and Oxford, 2019
18. Simon, H.A.: Motivational and emotional controls of cognition, *Psychological Review* 74:1 (1967) 29-39
19. Sloman, A., Croucher, M.: Why robots will have emotions, in: *IJCAI'81: Proc. 7th int. joint conf. on Artificial Intelligence*, Vol. 1, Morgan Kaufmann Publ., San Francisco, CA, 1981, pp. 197-202
20. Sowe, S.K., Simon, E., Zettsu, K., de Vaulx, F.F., Bojanova, I.: Cyber-Physical Human Systems: Putting People in the Loop, *IT Professional*, 18:1 (2016) 10-13
21. Synced: Michael I. Jordan Interview: Clarity of Thought on AI. In: *Synced*, November 29, 2018, https://syncedreview.com/2018/11/29/michael-i-jordan-interview-clarity-of-thought-on-ai/
22. Wiedermann, J., van Leeuwen, J.: Finite State Machines with Feedback: An Architecture Supporting Minimal Machine Consciousness. In: Manea F., Martin B., Paulusma D., Primiero G. (Eds): *Computability in Europe: Computing with Foresight and Industry* (CiE 2019), Lecture Notes in Computer Science, Vol. 11558, Springer, 2019, pp. 286-297
23. Wiedermann, J., van Leeuwen, J.: Towards minimally conscious finite-state controlled cyber-physical systems - a manifesto, *Technical Report* UU-PCS-2020-1, Center for Philosophy of Computer Science, Dept. of Information and Computing Sciences, Utrecht University, 2020.
24. Wikipedia: *Self-driving car*, March 2020, https://en.wikipedia.org/wiki/Self-driving_car
25. Zeng, J., Yang, L.T., Lin, M., Ning, H., Ma, J.: A survey: Cyber-physical-social systems and their system-level design methodology, *Future Generation Computer Systems* 105 (2020) 1028-1042