



Article

# Sharpening the Scythe of Technological Change: Socio-Technical Challenges of Autonomous and Adaptive Cyber-Physical Systems

Daniela Cancila <sup>1,\*</sup>, Jean-Louis Gerstenmayer <sup>2,†</sup>, Huascar Espinoza <sup>1</sup> and Roberto Passerone <sup>3</sup>

<sup>1</sup> CEA, LIST, CEA Saclay, PC172, 91191 Gif-sur-Yvette, France; huascar.espinoza@cea.fr

<sup>2</sup> MEFi-DGE French Ministry of Economy, 94201 Ivry-sur-Seine, France; jean-louis.gerstenmayer@cea.fr or jean-louis.gerstenmayer@finances.gouv.fr

<sup>3</sup> Dipartimento di Ingegneria e Scienza dell'Informazione, University of Trento, 38123 Trento, Italy; roberto.passerone@unitn.it

\* Correspondence: daniela.cancila@cea.fr; Tel.: +33-(0)1-6908-0107

† Researcher CEA at present in position of project manager and policy advisor at the French Ministry of Economy. Ideas and opinion in this paper are these of the author, they are not representative of the French Ministry of Economy opinions.

Received: 28 September 2018; Accepted: 21 November 2018; Published: 28 November 2018



**Abstract:** Autonomous and Adaptive Cyber-Physical Systems (ACPS) represent a new knowledge frontier of converging “nano-bio-info-cogno” technologies and applications. ACPS have the ability to integrate new ‘mutagenic’ technologies, i.e., technologies able to cause mutations in the society. Emerging approaches, such as artificial intelligence techniques and deep learning, enable exponential speedups for supporting increasingly higher levels of autonomy and self-adaptation. In spite of this disruptive landscape, however, deployment and broader adoption of ACPS in safety-critical scenarios remains challenging. In this paper, we address some challenges that are stretching the limits of ACPS safety engineering, including tightly related aspects such as ethics and resilience. We argue that a paradigm change is needed that includes the entire socio-technical aspects, including trustworthiness, responsibility, liability, as well as the ACPS ability to learn from past events, anticipate long-term threads and recover from unexpected behaviors.

**Keywords:** autonomous cyber-physical systems; resilience; ethics; nano-bio-info-cogno technologies

## 1. Introduction

In 2014, the Cyber-Physical European Roadmap and Strategy (CyPhERS) [1,2] pointed out artificial intelligence (AI) as a distinctive characteristic of cyber-physical systems (CPS). In 2018, after two years of research, the Platform4CPS European project [3] introduced a list of recommendations together with the main scientific topics and business opportunity for markets [4]. Among the main topics, the project highlights the importance of *autonomous Cyber-Physical Systems* (especially those including AI) and their impact on the incoming period with respect to several aspects and disciplines. For example, the introduction of AI into autonomous Cyber-Physical Systems demands new design and development technologies able to consider the learning phase, adaptability and requirement’s trace. Among the main concerns, the project highlights safety, resilience, security and confidence on these systems before their production and exploitation. Finally, the impact on social, legal and ethics issues is inherently involved by these systems. To clearly address the features of autonomy and environment adaptability of CPS [5], the European Commission refers to Autonomous Cyber-Physical Systems with the acronym ACPS. In this paper, we adopt then the term ACPS.

The design and development of ACPS requires the convergence of the cyber-side (computing and networking) with the physical side [6] and AI. More generally, tremendous progress becomes possible through converging technologies stimulated by advances in four core fields: *Nanotechnology, Biotechnology, Cognitive and Information technologies* [7,8]. Such convergence focuses on both the brain and the ambient socio-cultural environment. ACPS represent an example of that convergence which embraces not only engineering and technological products, but also legal (regulations) and ethical perspectives.

More specifically, the ACPS technologies are expected to bring large-scale improvements through new products and services across a myriad of applications ranging from healthcare to logistics through manufacturing, transport and more. The convergence of the “nano-bio-cogno-info” fields in ACPS could significantly improve the quality of our lives. However, the empowerment of sensitive intelligent components together with their increasing interaction with humans in shared spaces, e.g., personal care and collaborative industrial robots, raises pressing concerns about safety.

The technical foundations and assumptions on which traditional safety engineering principles are based worked well for human-in-the-loop systems where the human was in control, but are inadequate for ACPS, where autonomy and AI are progressively more active in this control loop. Incremental improvements in traditional safety engineering approaches over time have not converged to a suitable solution to engineer ACPS, having increased levels of autonomy and adaptation. In addition to physical integrity, safety in ACPS is tightly related to ethical and legal issues such as trustworthiness, responsibility, liability and privacy. A paradigm change is needed in the way we engineer and operate this kind of systems.

The main contribution of this paper is twofold: an analysis of the “nano-bio-cogno-info” convergence and a discussion on the aspects that are currently stretching the limits of ACPS safety engineering, including:

- (i) reduced ability to assess safety risks issues due to their analytical nature that relies on accident causality models,
- (ii) inherent inscrutability of Machine Learning algorithms due to our inability to collect an etymologically sufficient quantity of empirical data to ensure correctness,
- (iii) impact on certification issues, and
- (iv) complex and changing nature of the interaction with humans and the environment.

The results of both analysis highlight the importance of the notion of *resilience*. This notion emerges in the “nano-bio-cogno-info” convergence analysis as a means to the survival of the ‘competitors’ after an irreversible damage. Moreover, our analysis shows that traditional safety engineering cannot be applied to ACPS as it is. The European commission is aware of this topic and introduced it as high priority R&D&I area in the Strategic Research Agenda for Electronic Components and Systems [9]. Like the “nano-bio-cogno-info” convergence study, the emerging notion of resilience in the literature seems more appropriate than safety when we deal with ACPS. In [10,11], for example, the author highlights two ways to ensure safety: “avoiding that things go wrong” (called Safety I) and “ensuring that things go right” (named Safety II), i.e., the ability for a system to accomplish its mission under acceptable outcomes. In this regards -the author continues- “resilience does not argue for a replacement of Safety-I by Safety-II, but rather proposes a combination of the two ways of thinking”. But, more work should be done in this direction to establish safety engineering techniques and tools tailored to safety for ACPS.

The paper is organized as follows. Section 2 presents an overview of the nano-bio-cogno-info convergence. The section introduces the resilience issue and concludes with two open questions related to scientific and socio-ethics issues (Section 2.3) which are discussed in Section 3 and Section 4, respectively. More in particular, Section 3 discusses some scientific challenges to assess ACPS safety and Section 4 focuses on social and ethics issues, including privacy in ACPS. Finally, Section 5 shares our conclusions.

## 2. An Historical View of the Convergence Paradigm

The accelerating convergence of scientific and technological developments for the spheres of nano-systems industry, information and communication systems, as well as spheres of biology and medical applications, is an emerging phenomenon, based on our recently acquired capacity to measure, manipulate and organize the matter at the nanometer scale. A capital example of that convergence is the introduction of nano-material technologies tailored to medical applications. This is the case of intelligent microchip implants for humans (included in the brain) to control (lost) functionality. The design and development of that system demands microchips compatible with the human body, safety embedded software, prevention of potential attacks to the systems, and medical science. Quantum computing is another example: it combines neuromorphic architectures with computer science and physics. Finally, AlphaGo is able to automatically learn go, integrating artificial intelligence and automatic learning into embedded software and hardware [12].

Life sciences present the characteristics to become the keystone of that convergence. Nevertheless, as a consequence of maturing of artificial intelligence, more and more areas of convergence will be born ‘in silico’ before any ‘material’ development (e.g., scientists are building databases of thousands of compounds so that algorithms predict which ones combine to create new materials and design).

Historical Remarks: earlier digital technologies convergence and the development of interdisciplinary synergies are the precursors of a global (almost really ‘big’) convergence, known in the literature as Nanotechnologies (N), Biotechnologies (B), Information Technologies (I) and Cognitive sciences (C) (NBIC) convergence [7] or as Bio Nano & Converging Technologies (BNCT) [8]. In 1998, E. O. Wilson (sociobiologist) writes a book on the emerging harmony among the sciences [13] meaning the age of the concern of the knowledge unity (‘universe’ comes from ‘unus vs. multa’ (lat.)). However, the first consistent ‘Nano-Bio-Info-Cogno convergence’ paradigm was worldwide diffused only in 2005 by the very influential book of Mihail Roco and William Sims Bainbridge [7]. The authors prophesied a future overlapping of disciplines with extensive transdisciplinary fecundity. The figure represents the OCDE’s mapping, first introduced in [14], and reveals how distinct academic and technological domains can create an area of overlapping and increasing convergence (see Figure 1).

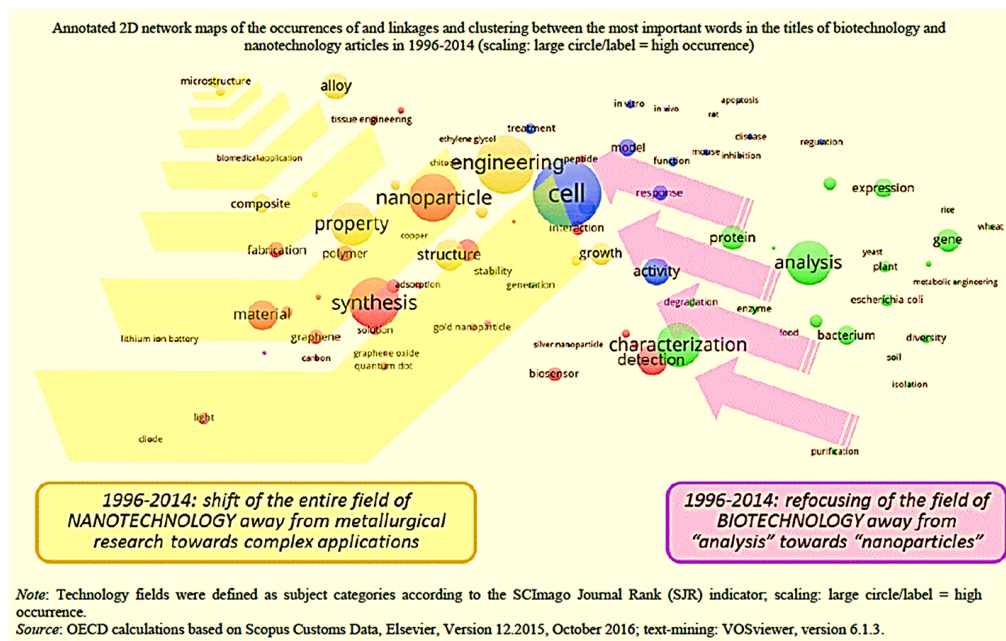


Figure 1. OCDE’s mapping on the convergence. The Figure is extracted from [14] with permission.

### 2.1. Several Underlying Realities in the Convergence

Several factors, scientific research, technological development, market adoption of diversified products and regulation, are contributing, differently, to convergence as a whole. A bird view may mask the distinction between the different aspects of the convergence having different motivations and distinct ethical concerns. The knowledge construction starts from non-resolved questions or theoretical dissonances. As achievement of our natural and permanent search for meaning and for efficiency, a theoretical convergence is first, regarding the theoretical sciences, the result of our natural need to know—when we synthesize powerful concepts unifying (anytime with genius) partial or multiple empirical knowledge in a global vision. As examples, when Einstein wrote, in 1905, his paper “Zur Elektrodynamik bewegter Körper” (On the Electrodynamics of Moving Bodies) known as ‘Einstein’s special theory of relativity’, he succeeded in the unification between Maxwell’s equations for electricity and magnetism with the laws of mechanics, which were formerly mutually inconsistent. In a second time, a technological background convergence results of our need for efficiency—when we build more complex patterns and combine means in order to fulfill new needs. The technological construction background starts from the lack of common functionality that has to be satisfied. We should not forget the recursive growth of sciences and technology: a progressive integration of previously independent disciplinary sciences leads to a larger disciplinary perimeter, increasing de facto the resource reservoir in use by engineers and inventors. Simultaneously, a cumulative growth of available technological solutions leads increasingly to a wider number of technological inventions potentially combinable. As a result, we obtain a combinatorial law of growth for the creative activity (obviously limited by a logistic function regarding the exchange fluxes or the management of stocks, on a longer period). Researchers propose diverse models to predict technological convergence and singularities [15]. Moreover, progress regarding the logistic limitations on the fluxes are also contributing to reinforce the acceleration and enhance the maximum level (e.g., digitalization, big data, human–machine interfacing, artificial intelligence) and the offer of capital determined by the level of trust. Truth is largely determined by governance and transparency issues. The judgment of rightness or wrongness upon the intellectual activity, regarding the knowledge construction, is the epistemology. The judgment on the goodness or the badness regarding an action, of individuals or of groups, having an impact on persons or on the whole society (equivalent to the Greek word *ethos*), is in the field of morals. As a whole, disruption, growth and convergence are largely driven by human appetite of knowledge and by trust. Trust is the fuel of business when science is the reliable rock where the imagination becomes rational. The concept of the technological convergence is becoming of greater importance because convergence is natural and its “road-map” highlights the decision making for more massive amount of funding needed by the industrial technology development. Then, enlarging the coherence on the scientific and technological basis ensures the investments and enlarges the opportunity of development. First benefit: to drive the convergence of micro-nano-bio technologies towards integrated biomedical systems. Imagination, initiative and boldness to launch new business are energies for good, but they always come with some indetermination and even with risks. Technological discoveries and innovations have always triggered fears and rejection, inspiring imagination for catastrophic and even apocalyptic scenarios. Recently, nanomaterials and nanotechnology incarnated the latest darkest fears related to technological development. Today, artificial intelligence with its increasing speed of development (relayed to the public for instance by the AlphaGo performance) is challenging the first place. The increasing speed of technological evolution, the enlarging perimeter of knowledge, the decreasing confidence in the expert, lead to a kind of cognitive dissonance due to the time required for constructing a coherent mental representation and a stable vision of our social ecosystem, thus, we both like and fear the technology, and the society is in tension: are we able to secure our trajectory of development? How could we proceed? This speed of change is often compared to a ‘shock’, a technological shock. Shock implies irreversibility, memory losses, change faster than the (social) speed of assimilation. The response to a change, to the indetermination, to a shock or to a probability of shock has to be discussed.

### 2.2. Resilience in the Convergence Paradigm

Innovation ability characterizes human societies and, anytime, fast developing social innovation or disruptive technological innovation feels as a shock for more vulnerable people. Physical sciences characterize a shock by a propagation speed larger than the speed of information (e.g., the speed of sound). Similarly, we can define a societal shock as a social transformation propagating faster than the ‘speed of acceptance’ along the social network (according to individual acceptance time and communication efficiency). A shock is also characterized by irreversibility, i.e., a partial ‘memory loss’ (explicitly coded -symbolic- information/or inherent -implicit- structural information (Structural information relative to a system, similarly to the philosophical notion of ‘essentia’, is from a bird view the ‘information needed to *build* it’, according to one of these two modalities of ‘being’: in knowledge (symbolic information) or with matter (inherent information) [16]) in the case of complex systems, and by a disruption of the physical state variables (temperature, pressure, speed of sound . . .). Resilience is the capability of a strained body to recover its size and shape after deformation caused especially by compressive stress [17]. In the field of societal issues, resilience is defined as “the ability to prepare and plan for, absorb, recover from, and more successfully adapt to adverse events” [18]. In the case of a learning autonomous-cyber-physical-system based on artificial intelligence, widely diffusing in the civil society [19], the capacity of *resilience* [from *resilire* (lat.), to back jump] seems to be a solution to restore this big ‘information system’, similar to—but more complex than—a software backup capacity. Nevertheless, it is not so simple when there are expensive infrastructures or humans in the loop and humans impacted. Hereafter, the ‘resilient’ subject, of interest, is holistically the whole interrelated system and some of its essential elements, including ACPS, humans (society), critical things and critical links. When is the resiliency most useful? According to Grime’s theory [20], ecological succession theory teaches us that plants adopt one survival strategy amongst three: (C) ‘Competitors’ maximize growth if resource is abundant and stable; (S) ‘Stress-tolerators’ maximize (individual) survival; (R) Ruderals maximize the survival of the species by means of a large fecundity. Two strategies are adapted to bad or very bad conditions of life, (S) to constantly unfavorable and (R) to largely fluctuating environment. Inversely, competitors (C) are optimized for growth in a context of resource abundance and are the least prepared for a (rare) shock (see Figure 2).

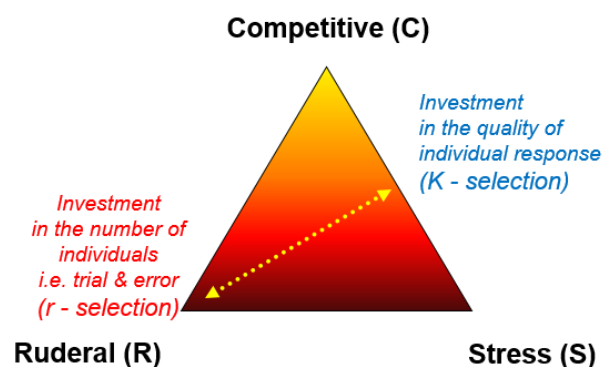


Figure 2. Illustration of the Philip Grime’s CSR theory.

As a conclusion, resilience is essential for the survival of the competitors after an irreversible damage leading from harmony to chaos (because of partial erasing of work and structural memory). Thus, what does resilience need? To be able to rebuild, resilience capacity needs, to be operational, an access to a preserved or a reconstructed memory (e.g., memory distributed in the social network) and an access to sources of efficiency (energy, tools, materials, money, willingness, etc.). Then the ability Resilience needs an efficient distribution of resources and of information about the system which can be stored outside of the system (strong importance of the weak links). In case of incomplete information, it may be difficult to rebuild the initial situation and the degraded resilience operation

may drive the system towards emerging new solutions of survival (evolution) worse or better adapted to the new conditions.

The case of France: with other countries, France participates in a conferences about “social impact” and Awareness-Readiness-Response [21]. The Joint Research Centre (JRC), as European Commission’s science and knowledge service, supports EU policies with independent scientific evidence throughout the whole policy cycle and develops innovative tools for policy makers challenges (including ethics, privacy, liability, etc.). In this context, JRC works on the social resilience too, defining a program of work (see Table 1). Basically, resilience or resiliency is the recovering of the initial state (resilire (lat.) = back jump) and ‘resistance’ is ‘remaining at the same place’.

**Table 1.** Information extracted from the table in [22].

What Could Be Tested?	How Should Be Tested?
Resilient Design	e.g., tests of the resiliency of a component or a system based on real failure data
(Inter)Dependency	Computer-based modeling and simulation (inter)dependency
Redundancy (including interoperability, adaptive capacity)	e.g., passive and active redundancy testing
Restoring capacity	e.g., measuring the ration of the lost performance

### 2.3. Open Questions

The technological convergence involves the notion of resilience, as described in the above section, and leaves two wide open questions:

- *A scientific issue* The restoration of a system at a previous state implies the use of memorized information. Thus, the resilience capacity is closely related to the accessibility to information on the oldest state to be restored, explicit and structural information, and to the accessibility to an energy source (physical energy, financial energy, social energy), leading to the obligation to preserve a good financial ratio between the costs of these functionalities (it may not be an economically sustainable option).
- *Socio-ethical questions* According to the point of view of the observer, the edge of a system is relative; it is simultaneously an ecosystem for smaller parts, a subsystem for a larger point of view, or the system itself for the designer. What part of the system must be protected: e.g., the ‘autonomous car’, its passengers, the dog in the street?

Life does not use invariably the resilience but often use the adaptability or the evolution, recovering a new equilibrium, different, even better than the oldest. Is it justifiable, on the economic or political view, to favor a back jump towards a past reference compared with other solutions more opened on an adaptive evolution?

In the next sections, we study these questions further.

### 3. Scientific Challenges of Safety-Critical ACPS

Autonomy in ACPS intrinsically involves an automatic decision-making and, then, more extensively, embraces Artificial Intelligence, as clearly stated in [1,2]. The introduction of AI in ACPS is revolutionizing safety techniques as traditionally used. Still a few year ago, some distinguished scientists wrote *Transcending Complacency on Superintelligent Machines*, acting as a gadfly to the community: “So, facing possible futures of incalculable [AI] benefits and [AI] risks, the experts are surely doing everything possible to ensure the best outcome, right? Wrong. [. . .] some of us—not only scientists, industrialists and generals—should ask ourselves what can we do now to improve the chances of reaping the [AI] benefits and avoiding the risks.” [23].

After that, in 2015, S. Russell et al. [24], in a letter signed by many scientists, point out three challenges related to AI safety:

- *verification* (how to prove that a system satisfies safety-related properties);
- *validation* (how to ensure that a system meets its formal requirements and does not have unwanted behaviors); and
- *control* (how to enable meaningful human control over an AI system after it begins to operate).

In 2016, the problems raised by safety and AI have been further analyzed in [25], where the authors focus on accidents in the machine learning systems. At the same time, the Future of Life Institute is becoming a leading actor for AI Safety Research [26].

ACPS are expected to bring a technological revolution that will influence the lives of a large part of the world population and will have a deep impact on nearly all market sectors. Increasing levels of AI will allow ACPS to drive on our roads, fly over our heads, move alongside us in our daily lives and work in our factories, offices and shops, soon. In spite of this disruptive landscape, deployment and broader adoption of ACPS in safety-critical scenarios remains challenging. Broadly speaking, the technical foundations and assumptions on which traditional safety engineering principles are based worked well for human-in-the-loop systems but are inadequate for autonomous systems where human beings are progressively ruled out from this control loop. The following sections discuss aspects that are currently stretching the limits of ACPS safety engineering. In other terms, they cannot be applied to ACPS as they are.

### 3.1. Ability to Appraise Safety Issues and to Learn from Experience

If we consider ACPS applications as a whole (e.g., drones, robots, autonomous vehicles), we discover that safety-related analysis, as applied in the traditional application domains such as railway or nuclear energy, suffers from given limits when we focus on ACPS. Of the extensive topic, we only discuss few examples that are, however, sufficiently expressive to show some safety-related limits.

Some of the most important assumptions in traditional safety engineering lie in models of how the world works. In this regards, current approaches include probabilistic risk assessment such as Preliminary Hazard Analysis, Fault Tree Analysis (FTA) and Failure Modes and Effects Analysis (FMEA), working on the assumption that, once the system is deployed, it does not learn and evolve anymore. Another strong hypothesis of the traditional safety approach is based on the constructor's responsibility. In traditional domains, specialized teams use the final product and are responsible of their maintenance. In the railway, for example, we have specialized train-drivers, teams specialized in the train maintenance, teams specialized in the physical infrastructure, team specialized in the electrical infrastructures, etc. The responsibility of an accident is totally in charge of the (train) constructors and/or the (railway) company. Rarely, a passenger is fully responsible to mitigate a possible accident. This situation does not happen in some applications of ACPS. Compare the railway organization with the case of a drone. The drone driver is not a safety engineer and/or a specialized drone pilot. Often, s/he has another work and uses the product, for example, to control agriculture or to film video. In case of an accident, the pilot's responsibility is analyzed. For example, in 2015, at the Pride Parade in Seattle, a drone's pilot has been judged guilty of a drone's accident and he was sentenced to 30 days in jail [27]. The role of a user of a (semi-autonomous) drone could be similar to the driver of a vehicle. Among the main differences, however, is the use of redundancy as a means to reduce the risk of an accident. In a vehicle, to ensure the precision of a measure, (a given set of) redundant and heterogeneous software and hardware mechanisms and architectures can be considered to be deployed. Of course, a vehicle is strongly limited in space and cost with respect to a train or a nuclear plant, but we can introduce some redundant mechanisms and architectures (e.g., more sensors to detect an obstacle). This technique, however, is strongly limited in the case of a drone due to space constraints. Preventing accidents in ACPS requires using models that include the entire socio-technical aspects and

treat safety as a dynamic control problem. Future intelligent autonomous systems need to be able to appraise safety issues in their environment, self-learn from experience and interactions with humans, and adapt and regulate their behavior appropriately. Furthermore, a higher level of autonomy in uncertain, unstructured, and dynamic environments involves many open systems science and systems engineering challenges. Autonomous systems interact in open-ended environments, making decisions and taking actions based on information from other systems and facing contingencies and events not known at design time. In recent years, new advances in technology have provided systems with the ability to anticipate, resist, recover from, and reconfigure after failures and disturbances. Safety assurance approaches have not kept up with the rapid pace of this kind of technological innovation.

#### Towards Ubiquitous Approaches for Resilience in ACPS

Recently, some scientists promoted methods and tools for *robust control* as a possible solution [28] to achieve the control of ACPS functionality even in the presence of uncertainty. Its cornerstone combines safety and performance over multi-sensor and multi-actuator heterogeneous networking architecture and—continue the authors in [28]—Model-Based Design represents a means to achieve a robust control design. Albeit robust control does not fix the AI-safety issues yet, it could be interesting to investigate in this direction further. First of all, robust control is closely associated with the concept of resilience [28]. We define two different types of resilience. *Exogenous* and *Endogenous* Resilience (see the side bar An Illustrative use case). Both types of resilience could represent a possible answer to ensure safety in ACPS. Hence, the paradigm changes: from safety, as traditionally studied in embedded systems, to resilience. Exogenous and Endogenous Resilience are pretty new in ACPS and, in our knowledge, established methodologies and tools for ACPS, including AI, do not exist yet.

An illustrative use case: drones are an illustrative example of ACPS and, extensively, of the technological convergence. Figure 3 shows the interaction between a fully-autonomous drone and a ground control station in the case of an accident. *Endogenous resilience* is related to the drone ability to accomplish its mission under physical constraints (e.g., battery level), error-detection (raising from embedded software, software and hardware integration, sensor failures, as well as malicious attack on the net), and safety-related measures to decrease the risk of an accident [29]. *Exogenous resilience* is related to the drone ability to accomplish its mission in presence of dynamic obstacles (e.g., birds) [29]. The Joint Authorities for Rulemaking of Unmanned Systems (JARUS) suggests some guidelines on Specific Operations Risk Assessment (SORA) for the category of specific drones [30], which do not require certification [31,32]. By the European regulation [31,32], only the category of fully automated drones ought to be certified. The blank hypothesis of the safety and certification process is that a drone does not change its internal behavior, for example, with respect to the number of flights (see Sections 3.1 and 3.3). This hypothesis is broken if AI is embedded in the software of the drone, for example by allowing a learning phase. Although specific drones do not require certification, the assessment of safety critical properties could be problematic by the introduction of AI and, in particular, of the machine learning algorithms, as discussed in [33].





**Figure 3.** Interaction between a fully-autonomous drone and a ground control station in the case of an accident. The figure is extracted from Emine Laarouchi. *An approach of safety analysis of CPS-IoT*. PhD on-going work. Supervisors: Daniela Cancila and Hakima Chaouchi.

### 3.2. Inherent Inscrutability of Autonomy Algorithms

Adaptation to the environment in autonomous systems is being achieved by AI techniques such as Machine Learning (ML) methods rather than more traditional engineering approaches. Recently, certain ML methods are proving themselves especially promising, such as deep learning, reinforcement learning and their combination. However, the inscrutability or opaqueness of the statistical models underlying ML algorithms (e.g., accidental correlations in training data or sensitivity to noise in perception and decision-making) pose yet another challenge. In traditional control systems, deductive inference logically links basic safety principles to implementation. Early cyber-physical systems tried to use deductive reasoning to construct sophisticated rule systems that fully defined behavior. However, cognitive systems have made spectacular gains using inductive inference based on ML, which may not produce semantically understandable rules, but rather find correlations and classification rules within training data. Inductive inference can yield excellent performance, in nominal conditions, but validating inductive learning is tough, due to our inability to collect an epistemologically sufficient quantity of empirical data to ensure correctness. The combination of autonomy and inscrutability in these inductive-based autonomous systems is particularly challenging for traditional safety techniques as their assurance is becoming intellectually unmanageable.

#### Towards Robustness Validation Approach for Autonomy

In the literature, one potential approach to deal with the inscrutability of AI-based systems is to identify the key elements of safety-related requirements on perception, see e.g., [34]. In that work, the authors specify the sources of perceptual uncertainty and suggest a reasoning on the system

(for example, Conceptual uncertainty, Development situation and scenario coverage, Situation or scenario uncertainty, etc.). Their idea is to control these factors to ensure that the system meets a threshold of acceptability.

Another potential approach to deal with the inscrutability of AI-based systems is robustness testing, which is a relatively mature technology for assessing the performance of a system under exceptional conditions. Fault Injection (FI) is a robustness testing technique that has been recognized as a potentially powerful technique for the safety assessment and corner-case validation of fault-tolerance mechanisms in autonomous systems [35]. The major aim of performing FI is not to validate functionality, but rather to probe how robust the vehicle is—or their components are—to arbitrary faults under unforeseen circumstances. The potential benefits of using FI into design phases of autonomous systems range from providing early opportunities for integration of inductive technologies—e.g., machine learning algorithms that use training sets to derive models of camera lens—to reducing costs and risks associated to autonomy functions. Such techniques have already been used successfully to find and characterize defects on autonomous vehicles [36].

### 3.3. Certification

The certification of a system happens after the design and development and the integration of the subsystems. A certified system can include subsystems having different level of criticality, i.e., different levels of certification [37].

Warning. Because of the description of the certification process in all details is extremely complex, we here provide only a simplistic overview of the process and we refer the interested reader to the related norms and literature.

Broadly speaking, a system is first specified by requirements, then designed, developed and, finally, (the components are) integrated. Safety analyses follow the design and development phases by ensuring compliance of the system under development to the referred safety-related standards (CENELEC for railway, ISO26262 for automotive, IEC 60880 for the nuclear, and so on). For example, in the railway application domain, safety teams assess the safety level (SIL) of the components during the initial phases of the design of the system, by proving the risk assessment analysis. During the system development, safety analyses exploit specific techniques to ensure that a component implements the SIL level, via, for example, the computation of the MTBF (mean time between failures) parameter and the implementation/justification of the underlying redundant system architecture (e.g., 2oo4). Once the system is integrated and analyzed, safety teams provide the arguments for the certification. A third certification entity first analyzes them and, then, discusses them via audit. We highlight that a system is always certified with respect to a particular use and related norms. For example, a system having a Technical Standard Order (TSO) authorization cannot be installed and used in an aircraft without passing the avionics certification [38].

Although the benefits of certified systems are marked and outstanding, the cost of the certification remains expensive and, in many case, prohibitive. In avionics, the cost to ensure the most critical level (level A) is estimated to be more than 55% with respect to the minimum level (level E) [37]. In the drone application domain, only one category requires certification ('Certified' Category for operations with an associated higher risk) [31,32].

#### Towards Certification Approaches for ACPS

The European commission has financially supported research projects in the last decade with the aim to reduce the cost of the certification. In this regard, the OPENCOSS European project integrates the certification's arguments already in the preliminary phases of the design [39]; Goal Structuring Notation (GSN) is a graphical formalism to safety arguments [40].

Today, certification studies mechanisms to ensure modularity and mixed-criticality on many and multi-core. Mixed-criticality is the ability of a system to execute functionality, having different levels of criticality, in the same hardware by guaranteeing the safety level associated to each function.

The PROXIMA European project [41] is part of the mixed-criticality European cluster, together with CONTREX [42] and DREAMS [43] European projects, financially supported by the European commission. The PROXIMA project analyzes “software timing using probabilistic analysis for many and multi-core critical real-time embedded systems and will enable cost-effective verification of software timing analysis including worst case execution time” [41]. The result is compliant with DO-178B (Software Considerations in Airborne Systems and Equipment Certification). A modular certification aims to study mechanisms to restrain certification from the certified system as a unique whole to the certified component, which is modified or substituted. These mechanisms primarily address contract-based system interfaces [44], and impact analysis of the component modification/substitution in the remaining system. One main difficulty of the certification process does not concern correctness of the automatically generated code with respect to a given model, but the assurance that the system, which is executed, meets the system requirements and only implements the specified functionality. In other terms, engineers have to avoid to automatically generate correct code with respect to a wrong model specification. As in the case of safety, certification is based on the hypothesis that the system to be certified will have the same behavior (it always implements only the specified functionality). In other words, no learning phase is allowed. For example, in the avionic application domain, a program does not change its behavior with the increasing number of performed flights. Similarly, in the railway application domain, the control command to automatically open/close the doors of an automatic metro will have the same behavior forever. However, the AI introduction and, in particular, the (on-line or off-line) learning phase overstretch the fundamental hypothesis on which the traditional certification process is based (i.e., the learning phases overstretch reproducibility of proofs that ensure the same system behavior under the same inputs). This situation is also problematic for the category of specific drones (which do not require certification [31,32]) to assess safety-related requirements [30] as discussed in [33].

In addition, if AI is coupled with a full or a high level of system autonomy (i.e., human does not have control of the system), then they unwind civilian responsibility: who is responsible of an accident of ACPS having AI? How can we protect the ACPS and humans? Is it possible to certify an ACPS having AI? And how?

In this regards, an interesting study is *Certification Considerations for Adaptive Systems* by NASA [45]. In the document, the authors address the impact on adaptive and AI system in the avionics certification. The authors conclude with a number of recommendations and a road-map to improve certification approaches. One issue is to relax some strict assumptions and requirements to allow a more streamlined certification process.

### 3.4. New Forms of Interaction between Humans and Autonomous Systems

There are unique challenges in moving from human-machine interaction in automation, where machines are essentially used as tools, to the complex interactions between humans and autonomous systems or agents. As autonomous systems progressively substitute cognitive human tasks, some kind of issues become more critical, such as the loss of situation awareness, or the overconfidence in highly-automated machines. The Tesla accident occurred in 2016 is a clear illustration of the loss of situation awareness in semi-autonomous vehicles, as stated by the National Transportation Safety Board (NTSB): “the cause of the crash was overreliance on automation, lack of engagement by the driver, and inattention to the roadway” [46]. One of the main causes of this problem is a reduced level of cognitive engagement when the human becomes a passive processor rather than an active processor of information [47].

In [48], the author addresses the “Ironies of Automation”, especially for the control process in industries. The main paradox is that any automated process needs (human) supervisors. In the case of an accident or a problem in industry, the human supervisor may not be sufficiently prepared or reactive to solve it, because automation hides and directly manages ex-manual operations. Therefore, the supervisor could be less prepared in autonomous industrial control systems. Of course,

this situation could appear in industry and not in traditional critical systems. In a nuclear plant, for example, engineers in the control rooms receive an expensive and strong (theoretical and experimental via simulation) training, required after the TMI-2 accident [49,50].

While the potential loss of situation awareness is particularly relevant in autonomous systems needing successful intervention of humans, there is a number of more general situations where risks in human-machine interaction must be better understood and mitigated:

- Collaborative missions that need unambiguous communication (including timescale for action) to manage self-initiative to start or transfer tasks.
- Safety-critical situations in which earning and maintaining trust is essential at operational phases (situations that cannot be validated in advance). If humans determine the system might be incapable of performing a dangerous job, they would take control of the system.
- Cooperative human-machine decision tasks where understanding machine decisions are crucial to validate autonomous actions. This kind of scenario implies providing autonomous agents with transparent and explainable cognitive capabilities.

#### Towards Trusted and Safe Human-Machine Relationships

Several authors [51–54] argue that interactions with autonomous agents must be considered as “human relationships”, as we are delegating cognitive tasks to these entities. This perspective opens the door to the application of existing fundamental knowledge from the social sciences (psychology, cognitive modelling, neuropsychology, among others), to develop trustable and safe interactions with autonomous systems. For instance, the authors of [51] propose to encode the human ability of building and repairing trust into the algorithms of autonomous systems. They consider trust repair a form of social resilience where an intelligent agent recognises its own faults and establishes a regulation act or corrective action to avoid dangerous situations. Unlike other approaches where machine errors remain unacknowledged, this approach builds on creating stronger collaboration relationships between humans and machines to rapidly adjust any potential unintended situation.

In 2017, some influential leaders in AI established the Asilomar AI Principles aimed at ensuring that AI remains beneficial to humans [55]. One of these principles is value alignment, which demands that autonomous systems “should be designed so that their goals and behaviours can be assured to align with human values throughout their operation”. In this context, some researchers such as Sarma et al. [53] argue that autonomous agents should infer human values by emulating our social behaviour, instead of embedding these values into their algorithms. This approach would also apply to the way human interact and it would be the basis to create learning mechanisms emphasizing trust and safe behaviours, including abilities such as intentionality, joint attention, and behaviour regulation. While these ideas look more appealing with the rise of AI and machine learning, the concept of learning “safe” behaviours in intelligent agents was already explored in 2002 by Bryson et al. [54]. Nevertheless, the problem of finding meaningful safety mechanisms for human-machine interaction inspired from human social skills remains largely open, because of the complex and intricate nature of human behaviour and the need of a provably-safe framework to understand and deploy artificial relationships.

#### 4. Socio-Ethical Challenges of Safety-Critical ACPS

Social trust is a major challenge for the future of ACPS. The recent catastrophic accidents involving autonomous systems (e.g., Tesla fatal car accident), show that sole engineering progress in the technology is not enough to guarantee a safe and productive partnership between a human and an autonomous system. The immediate technical research questions that come to mind are how to quantify social trust and how to model its evolution? Another direction that is key to understanding and formalizing of social trust is how to design a logic that allows the expression of specifications involving social trust? It is immediately followed by the questions of how to verify (reason about) such specifications in the context of a given human-machine collaborative context or, even more prominent,

how to synthesize (design) an autonomous system such that, in a collaborative context with a human, these specifications are guaranteed?

#### 4.1. Ethics and Liability in ACPS

Ethics in safety-critical autonomous systems is closely related to the question of risk transfer. If any safety risk is transferred from some people to others then the risk transfer must be explicitly justified, even where the overall risk is reduced. Indeed, while it may be possible to argue that the introduction of ACPS in certain situations (e.g., automated cars) reduces the overall harm, from an ethical point of view this may not be sufficient. Ethical issues can be regarded in terms of the trade-off associated with reducing one risk posed by an ACPS at the potential cost of increasing another risk. This is essential to understand ethics principles in terms of safety [56]. In the literature, Andreas Matthias introduced the notion of *responsibility gaps* to identify the situation in which “nobody has enough control over the machine’s actions to assume the responsibility for them” [57]. This notion has been developed further to cover two dimensions: the control on the “what” and on the “how” of the system behaviors [58].

One fundamental problem when dealing with ACPS is the liability for accidents involving autonomous systems. As a general rule, the more a machine can learn, the less control the manufacturer has, which is important for product liability. On the 16 of February 2017, following the suggestion made by the Legal Affairs Committee (LAC), the European Parliament (EP) made a resolution in favor of a robust European legal framework to ensure that autonomous systems are and will remain in the service of humans. Regarding liability, the EP suggests to explore, analyze, and consider the implications of all possible legal solutions, such as “establishing a compulsory insurance scheme”, “compensation funds” or even “creating a specific legal status for autonomous systems in the long run”. Other studies suggest addressing ACPS-specific risks in comparison with the risks of traditional technologies (without learning or adaptive abilities) used for the same purpose. This may facilitate the analysis of liability cases under existing law or be used as elements in new legal rules.

An important issue is related to the implementation of ethical and other legal rules in the inherent behavior of autonomous systems. In this area we need:

- Support the modeling, verification and transformation of safety and ethical rules into machine-usable representations (safety/ethics constraint set) and reasoning capabilities for run-time monitoring and adaptation. This includes the integration of ethics and safety in the whole system engineering life-cycle.
- Embody building blocks (regulators, adviser, adaptors) to ensure that the design of intelligent behaviors only provide responses within rigorously defined safety and ethical boundaries (safety/ethics constraint set). This includes the definition of architectural patterns to systematically reduce the development complexity -including integration- and modular assurance of AI-based systems.
- Support mechanisms to permit the monitoring and adaptation of a safety constraint set and underlying behavioral control parameters. Human-machine interaction abilities to warn human operators and users about safety issues preventing the autonomous system from acting, to manage the override of a constraint, and to inform about any change in the behavioral (safety/ethical) constraint set.

#### 4.2. Privacy

Privacy is a live issue in the society and crosses several levels. In [4] privacy is identified as a societal and legal grand challenge. It is especially emphasized whenever AI functionality will be widely developed in the ACPS systems. In [24], for example, the authors reveal how “the ability of AI systems to interpret the data obtained from surveillance cameras, phone lines, emails, and so on, can interact with the right to privacy”. In this section we investigate the issue via two examples: one related to drones and one to autonomous vehicles. These examples show how the privacy right has

acquired importance in our society and could be weakened by technical solutions on ACPS. We share the Platform4CPS vision and recommendation [4] which states that privacy will be a major topic in the future, reinforced by the AI introduction and AI expected exploitation on ACPS products.

#### 4.2.1. Privacy in the Use of Drones

In 2015, just after an extremely violent earthquake, Nepal decided to use drone technology. These later have proved to be extremely useful in catastrophic scenarios. Drones are expected to increase the chance of survival to 80% because they are more reactive and, then, more efficient than the traditional surgery means [59]. For example, drones have been used after the Irma Hurricane in Florida (2017) to understand the situation and target the first-aid, and in Paris (Spring 2018) to control the dangerous increasing level of Senna and prevent catastrophic consequences for humans and the city (Several videos and resources are available on the net. For example, Paris [https://www.francetvinfo.fr/meteo/inondations/video-la-crue-de-la-seine-filmee-par-un-drone-a-paris\\_2580526.html](https://www.francetvinfo.fr/meteo/inondations/video-la-crue-de-la-seine-filmee-par-un-drone-a-paris_2580526.html) and Florida <https://unmanned-aerial.com/drones-artificial-intelligence-impacted-hurricane-irma-claims-response>). However, in Nepal drones have been banned after their use in the earthquake, because of a privacy conflict due to having captured images on heritage sites [60]. Privacy issue for drone applications is an extremely sensitive issue, as analyzed by the works of Silvana Pedrozo and Francisco Klausner (sociologists). The authors are (also) known to having made a *sondage* on the Swiss population on the acceptability or not to drones in 2015. The results have been diffused and discussed via several means (radio, scientific journals, etc.) [61,62]. In [61], the authors state that: “whilst the majority of respondents are supportive of the use of unarmed military and police drones (65 and 72% respectively), relative numbers of approval decrease to 23 and 32% when it comes to commercial and hobby drones”. The underlying reason of such acceptability is not based on the guarantee of safety level of that systems. For example, in 2015, a drone injured a woman at the Pride Parade in Seattle [27] and the news has had a very broad resonance [63]. Instead, it is based on privacy concern and individual freedom [62]. This challenge is so relevant that the European regulation for drones addresses privacy issue with the support of a lawyer European team [64] (see, for instance, Article 29 Data Protection Working Party). Similarly, the Platform4CPS project [3] suggests to “enforce General Data Protection Regulation (GDPR), mandate in-built security mechanisms for key applications and clarify liability law for new products and services”. The project identifies the following potential implementation: “Put in place enforcement measures for GDPR, enforce built in security for European products and put in place appropriate legislation for products and services” [4].

#### 4.2.2. Privacy in Autonomous Vehicles Communications

Privacy covers more aspects in autonomous vehicles: from embedded software (e.g., how we deal with the personal data of the driver) to vehicle to vehicle (V2V) communication. Of the extensive issue, we only focus here on the latter.

The main problem concerning privacy in the V2V communications regards the full traceability of a vehicle. Vehicles should communicate with each other, for example to signal an accident or know if another vehicle is incoming to the same crossroad. The simplest technical solution to manage this scenario is to provide a unique identification to each vehicle. Problem solved. However, this technical solution collides with privacy: everyone can trace the trajectory of sensible targets. Examples of sensible targets are people having a protection, transfer of jailed people, bank transfer. All these targets can be easily traced and then, potentially, attacked. Moreover, this social phenomenon could have a wider impact: a wife/husband can trace the partner, an employer traces the employees, etc. To avoid a scenario, as described in the book 1984 by George Orwell [65], the technical solution is a pseudonym for each vehicle. However, the social impact is similar to the previous one: it is technically a simple game to understand the association pseudonym/target, trace it and attack it. Another technical solution consists in periodically changing the pseudonym. However, if this change is too slow, Bob can easily trace Alice. On the contrary, if it is too fast, an autonomous vehicle receives the information that

ten vehicles are incoming in the same crossroad whilst, in fact, only one is engaging the crossroad (the vehicle has automatically changed the pseudonym nine times in a short elapsed time).

This example clearly shows how technical solutions impact our society and could limit some rights, as privacy. This technical and society challenge is extremely sensible in the community: a special working group in ETSI TC ITS is addressing the question [66,67], the European commission is devoting its financial effort to support several projects on that direction (e.g., see [68]). Finally, in the literature we can find interesting surveys regarding privacy and V2V communications (see for example [69]), technical investigations and promoted solutions (see e.g., [70]).

## 5. Conclusions

We are witnessing a convergence in Nanotechnologies, Biotechnologies, Information technologies and Cognitive sciences [7]. In this regard, our analysis showed the importance of the notion of resilience (Section 2.2) and opens the door to two open questions on scientific and socio-ethical issues (Section 2.3). The rest of the paper developed both questions further. To do this, we focus on ACPS which are a key example of the convergence in Nanotechnologies, Biotechnologies, Information technologies and Cognitive sciences convergence and we mainly addressed four topics: (i) reduced ability to assess safety risks, (ii) inherent inscrutability of Machine Learning algorithms, (iii) impact on certification issues and (iv) complex and changing nature of the interaction with humans and the environment.

(i) and (iii) ACPS having AI and a high level of autonomy are showing the limits of safety and certification processes when applied to these systems. Traditional safety techniques and related methods cannot be applied as they are. (ii) The core concern is the unpredictability due to machine learning, which is, on one hand, a means to reduce and manage complexity inherently involved by ACPS, but, on the other hand, there still are no widely accepted techniques and processes to manage the impact of machine learning in the assessment of safety-related properties and certification. Hereafter, the main response to this kind of issue are collected, where “responsibility sharing” is the key concept in addition to other levels of response (resilience, autopoiesis, insurance, legal solution, etc.). As a generic example, AI inside ACPS leads to new issues due to unpredictability of the comportment emerging from unpredicted competences and behavior that machine learning will develop in different contexts of use. This unpredictability may be caused by the impossibility for the designers and safety teams to specify the multiplicity of situations and contexts, or by the practical impossibility to study a behavior that will be learned during the life of the system, or by the practical impossibility to compute them, or by the practical impossibility to collect all return experience data of interest. Hence, there are too many cases or these cases are difficult to be defined precisely, so that it is difficult to build an exhaustive base of examples dedicated to constructing exhaustive incidental cases virtual simulations as a base of reliability. (iv) In addition, an inadequate communication between humans and machines is becoming an increasingly important factor in accidents, which can create significant new challenges in the areas of human–machine interaction and mixed initiative control.

One interesting solution, proposed by NASA [45], is to limit the range of possibilities and relax some hypotheses by adopting a more streamlined (safety and certification) process. A similar problem appears in the education of a young human: some of them may become incompetent or immoral. Nevertheless, despite this individual human unpredictability, a human society is globally stable on the long term because people have individually and collectively the ability to learn a safer behavior (cautiousness) and to become ‘responsible’. What can we do in the case of a hybrid society including learning machines? It is possible to transpose, adapt or divide the notion of ‘responsibility’ (technical and legal point of view), alone or with the complement of other dispositions as insurance and resilience? These are the way of progress to be explored.

**Author Contributions:** J.-L.G. contributed to Sections 2, 2.1, 2.2, 2.3 and 5. H.E. contributed to Sections 1, 3, 3.1, 3.2, 3.4 and 4.1. D.C. contributed to Sections 1, 3, 3.1, 3.3, 4.2 and 5, LaTeX Editing and team coordination. R.P. has reviewed all the work.

**Acknowledgments:** We thank Françoise Roure of Ministère de l'Économie et des Finances—France for her studies and suggestions on social resilience. We thank Steffi Friedrichs for the figure “OECD’s mapping”. We thank Emine Laarouchi for the figure with drones and for his help with the reviews integration. We thank the journal editorial board and the reviewers for their comments and valuable suggestions.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ACPS	Autonomous and Adaptive Cyber-Physical Systems
BNCT	Bio Nano & Converging Technologies
EP	European Parliament
ETSI TC ITS	ETSI Technical Committee (TC) for Intelligent Transport Systems (ITS)
FI	Fault Injection
ML	Machine Learning
NBIC	Nano-Bio-Info-Cogno (technologies)
V&V	vehicle to vehicle

## References

1. CyPhERS FP7 Project. Cyber-Physical European Roadmap and Strategy. Available online: <http://cyphers.eu/> (accessed on 26 November 2018).
2. Törngren, M.; Asplund, F.; Bensalem, S.; McDermid, J.; Passerone, R.; Pfeifer, H.; Sangiovanni-Vincentelli, A.; Schätz, B. Characterization, Analysis, and Recommendations for Exploiting the Opportunities of Cyber-Physical Systems. In *Cyber-Physical Systems*; Song, H., Rawat, D.B., Jeschke, S., Brecher, C., Eds.; Intelligent Data Centric Systems; Academic Press, Elsevier: Cambridge, MA, USA, 2016; Chapter 1, pp. 3–14.
3. Platform4CPS European Project. Available online: <https://www.platforms4cps.eu/> (accessed on 26 November 2018).
4. Thompson, H.; Reimann, M.; Ramos-Hernandez, D.; Bageritz, S.; Brunet, A.; Robinson, C.; Sautter, B.; Linzbach, J.; Pfeifer, H.; Aravantinos, V.; et al. *Platforms4CPS: Key Outcomes and Recommendations*; Steinbeis: Stuttgart, Germany, 2018.
5. D’Elia, S. *CPS in EU Programmes*; European Commission DG CONNECT: Brussels, Belgium, 2017.
6. Sangiovanni-Vincentelli, A.; Damm, W.; Passerone, R. Taming Dr. Frankenstein: Contract-Based Design for Cyber-Physical Systems. *Eur. J. Control* **2012**, *18*, 217–238. [[CrossRef](#)]
7. Bainbridge, W.S.; Roco, M.C. (Eds.) *Managing Nano-Bio-Info-Cogno Innovations: Coverging Technologies in Society*; Springer: New York, NY, USA, 2005.
8. Winickoff, D. *Working Party on Biotechnology, Nanotechnology and Converging Technologies: BNCT Project Updates*; Technical Report DSTI/STP/BNCT(2015)6; OECD: Paris, France, 2015.
9. Aeneas. *Strategic Research Agenda for Electronic Components and Systems (ECS-SRA)*; EpoSS: Berlin, Germany, 2018.
10. Woods, D.D.; Hollnagel, E. *Resilience Engineering: Concepts and Precepts*; CRC Press: Boca Raton, FL, USA, 2006.
11. Hollnagel, E. *Safety-I and Safety-II: The Past and Future of Safety Management*; CRC Press: Boca Raton, FL, USA, 2014.
12. Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. Mastering the game of Go without human knowledge. *Nature* **2017**, *550*, 354–359. [[CrossRef](#)] [[PubMed](#)]
13. Wilson, E.O. *Consilience: The Unit of Knowledge*; Alfred A. Knopf: New York, NY, USA, 1998.
14. Friedrichs, S. *Report on Statistics and Indicators of Biotechnology and Nanotechnology. Documents de Travail de L’OCDE sur la Science, la Technologie et L’Industrie*; Technical Report 06; OECD: Paris, France, 2018.
15. Sandberg, A. An Overview of Models of Technological Singularity. In *The Transhumanist Reader*; Wiley-Blackwell: Hoboken, NJ, USA, 2013; Chapter 36, pp. 376–394.



16. Burgin, M.; Feistel, R. Structural and Symbolic Information in the Context of the General Theory of Information. *Information* **2017**, *8*, 139. [CrossRef]
17. Merriam-Webster. Dictionary. Available online: <https://www.merriam-webster.com/dictionary/resilience> (accessed on 26 November 2018).
18. The National Academies Press. *Disaster Resilience: A National Imperative*; The National Academies Press: Washington, DC, USA, 2012.
19. Villani, C. Donner un Sens à l'Intelligence Artificielle. Rapport au Premier Ministre, Mission Confiée par le Premier Ministre Edouard Philippe, Mission Parlementaire du 8 Septembre 2017 au 8 Mars 2018. Available online: <https://www.ladocumentationfrancaise.fr/var/storage/rapports-publics/184000159.pdf> (accessed on 26 November 2018).
20. Silvertown, J.; Franco, M.; McConway, K. A Demographic Interpretation of Grime's Triangle. *Funct. Ecol.* **1992**, *6*, 130–136. [CrossRef]
21. Roure, F. Nanotechnology: 10 Years of French Public Policy towards a Responsible Development. Available online: [https://www.youtube.com/watch?v=W4QzHaok\\_Xo](https://www.youtube.com/watch?v=W4QzHaok_Xo) (accessed on 26 November 2018).
22. Pursiainen, C.; Gattinesi, P. *Towards Testing Critical Infrastructure Resilience*; Joint Research Centre (JRC) Scientific and Policy Reports; JRC: Ispra, Italy, 2014.
23. Huffington Post. Stephen Hawking and Max Tegmark and Stuart Russell and Frank Wilczek. In *Transcending Complacency on Superintelligent Machines*; Huffington Post: New York, NY, USA, 2014.
24. Russell, S.J.; Dewey, D.; Tegmark, M. Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Mag.* **2015**, *36*, 105–114. [CrossRef]
25. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.F.; Schulman, J.; Mané, D. Concrete Problems in AI Safety. *arXiv* **2016**, arXiv:1606.06565.
26. Future of Life. AI Safety Research. Available online: <https://futureoflife.org/ai-safety-research/> (accessed on 26 November 2018).
27. Steve Miletich. Pilot of Drone That Struck Woman at Pride Parade Gets 30 Days In Jail. 2017. Available online: <https://www.seattletimes.com/seattle-news/crime/pilot-of-drone-that-struck-woman-at-pride-parade-sentenced-to-30-days-in-jail/> (accessed on 26 November 2018).
28. Amin, S.; Schwartz, G.; Sastry, S.S. *Challenges for Control Research: Resilient Cyber-Physical Systems*; Technical Report; IEEE CSS: New York, NY, USA, 2014.
29. Laarouchi, E.; Mouelhi, S.; Cancila, D.; Chaouchi, H. Robust Control Predictive Design for Resilient Cyber-Physical Systems. *Ada Eur.* **2019**, submitted.
30. Joint Authorities for Rulemaking of Unmanned Systems (JARUS). *JARUS Guidelines on Specific Operations Risk Assessment (SORA)*; Technical Report JAR-DEL-WG6-D.04; Swiss Federal Office of Civil Aviation (OFAC): Ittigen, Switzerland, 2017.
31. European Aviation Safety Agency. *Advance Notice of Proposed Amendment (A-NPA) 2015-10—Introduction of a Regulatory Framework for the Operation of Drones*; EASA: Brussels, Belgium, 2015.
32. European Aviation Safety Agency. 'Prototype' Commission Regulation on Unmanned Aircraft Operations; EASA: Brussels, Belgium, 2016.
33. Schirmer, S.; Torens, C.; Nikodem, F.; Dauer, J. Considerations of Artificial Intelligence Safety Engineering for Unmanned Aircraft. In Proceedings of the First International Workshop on Artificial Intelligence Safety Engineering (WAISE), Vasteras, Sweden, 18 September 2018.
34. Czarnecki, K.; Salay, R. Towards a Framework to Manage Perceptual Uncertainty for Safe Automated Driving. In Proceedings of the First International Workshop on Artificial Intelligence Safety Engineering (WAISE), Vasteras, Sweden, 18 September 2018.
35. Juez, G.; Amparan, E.; Ruiz, A.; Perez, J.; Lattarulo, R.; Espinoza, H. Early Safety Assessment of Automotive Systems Using Sabotage Simulation-Based Fault Injection Framework. In *Computer Safety, Reliability, and Security, Proceedings of the International Conference on Computer Safety, Reliability, and Security, Trento, Italy, 12–15 September 2017*; Springer: Cham, Switzerland, 2017.
36. Vernaza, P.; Guttendorf, D.; Wagner, M.; Koopman, P. Learning Product Set Models of Fault Triggers in High-Dimensional Software Interfaces. In Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October 2015.
37. Hilderman, V.; Baghai, T. *Avionics Certification: A Complete Guide to DO-178 (Software) DO-254 (Hardware)*; Avionics Communicaitons Inc.: Leesburg, VA, USA, 2008.

38. FAA. Aircraft Certification Design Approvals Technical Standard Order (TSO). Available online: [https://www.faa.gov/aircraft/air\\_cert/design\\_approvals/tso/](https://www.faa.gov/aircraft/air_cert/design_approvals/tso/) (accessed on 26 November 2018).
39. OPENCROSS Open Platform for Evolutionary Certification of Safety-Critical Systems (OPENCROSS), FP7 European Project. Available online: <http://www.opencross-project.eu/> (accessed on 26 November 2018).
40. The GSN Working Group Online. Goal Structuring Notation (GSN). Available online: [http://www.goalstructuringnotation.info/documents/GSN\\_Standard.pdf](http://www.goalstructuringnotation.info/documents/GSN_Standard.pdf) (accessed on 26 November 2018).
41. PROXIMA FP7 European Project. Probabilistic Real-Time Control of Mixed-Criticality Multicore and Manycore Systems (PROXIMA). Available online: [https://cordis.europa.eu/project/rcn/109947\\_fr.html](https://cordis.europa.eu/project/rcn/109947_fr.html) (accessed on 26 November 2018).
42. CONTREX FP7 European Project. Design of Embedded Mixed-Criticality CONTROL Systems under Consideration of EXtra-Functional Properties (CONTREX). Available online: <https://contrex.offis.de/home/> (accessed on 26 November 2018).
43. DREAMS FP7 European Project. Distributed REal-Time Architecture for Mixed Criticality Systems (DREAMS), FP7 European Project. Available online: <https://contrex.offis.de/home/> (accessed on 26 November 2018).
44. Benveniste, A.; Caillaud, B.; Nickovic, D.; Passerone, R.; Raclet, J.B.; Reinkemeier, P.; Sangiovanni-Vincentelli, A.L.; Damm, W.; Henzinger, T.A.; Larsen, K.G. Contracts for system design. *Found. Trends Electron. Des. Autom.* **2018**, *12*, 124–400. [CrossRef]
45. Bhattacharyya, S.; Cofer, D.; Musliner, D.J.; Mueller, J.; Engstrom, E. *Certification Considerations for Adaptive Systems*; Technical Report NASA/CR-2015-218702; NASA: Washington, DC, USA, 2015.
46. National Transportation Safety Board. *Collision between a Car Operating with Automated Vehicle Control Systems and a Tractor-Semitractor Truck near Williston, Florida, 7 May 2016*; Accident Report NTSB/HAR-17/02-PB2017-102600; National Transportation Safety Board: Washington, DC, USA, 2017; p. 42.
47. Endsley, M.R. Situation Awareness in Future Autonomous Vehicles: Beware of the Unexpected. In *Advances in Intelligent Systems and Computing, Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018), Florence, Italy, 26–30 August 2018*; Bagnara, S., Tartaglia, R., Albolino, S., Alexander, T., Fujita, Y., Eds.; Springer: Cham, Switzerland, 2018; Volume 824.
48. Bainbridge, L. Ironies of Automation. *Sci. Direct* **1983**, *19*, 775–779.
49. Clément, B.; Jacquemain, D. *Nuclear Power Reactor Core Melt Accidents*; Chapter Lessons Learned from the Three Mile Island and Chernobyl Accidents and from the Phebus FP Research Programme—Chapter 7; IRSN: Paris, France, 2015.
50. Walker, S. *Three Mile Island a Nuclear Crisis in Historical Perspective*; University of California Press: Berkeley, CA, USA, 2006.
51. de Visser, E.J.; Pak, R.; Shaw, T.H. From automation to autonomy: The importance of trust repair in human-machine interaction. *J. Ergon.* **2018**. [CrossRef] [PubMed]
52. Kohn, S.C.; Quinn, D.; Pak, R.; de Visser, E.J.; Shaw, T.H. *Trust Repair Strategies with Self-Driving Vehicles: An Exploratory Study*; SAGE Publications: Thousand Oaks, CA, USA, 2018; Volume 62, pp. 1108–1112.
53. Sarma, G.P.; Hay, N.J.; Safron, A. AI Safety and Reproducibility: Establishing Robust Foundations for the Neuropsychology of Human Values. In *Computer Safety, Reliability, and Security*; Gallina, B., Skavhaug, A., Schoitsch, E., Bitsch, F., Eds.; Springer: Berlin, Germany, 2018; pp. 507–512.
54. Bryson, J.J.; Hauser, M.D. What Monkeys See and Don't Do: Agent Models of Safe Learning in Primates. In *Proceedings of the AAAI Symposium on Safe Learning Agents*, Palo Alto, CA, USA, 25–27 March 2002.
55. The Future of Life Institute. Asimolar AI Principles. Available online: <https://futureoflife.org/ai-principles/> (accessed on 1 November 2018).
56. Menon, C.; Alexander, R. Ethics and the safety of autonomous systems. In *Proceedings of the Safety-Critical Systems Symposium*, York, UK, 6–8 February 2018.
57. Matthias, A. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics Inf. Technol.* **2004**, *6*, 175–183. [CrossRef]
58. Porter, Z.; Habli, I.; Monkhouse, H.; Bragg, J. The Moral Responsibility Gap and the Increasing Autonomy of Systems. In *Proceedings of the First International Workshop on Artificial Intelligence Safety Engineering (WAISE)*, Vasteras, Sweden, 18 September 2018.
59. Scott, J.E.; Scott, C.H. Drone Delivery Models for Healthcare. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*, Honolulu, HI, USA, 4–7 January 2017.

60. Choi-Fitzpatrick, A.; Chavarria, D.; Cychosz, E.; Dingens, J.P.; Duffey, M.; Koebel, K.; Siriphanh, S.; Yurika Tulen, M.; Watanabe, H.; Juskauskas, T.; et al. *Up in the Air: A Global Estimate of Non-Violent Drone Use 2009–2015*; Technical Report; University of San Diego Digital USD: San Diego, CA, USA, 2016.
61. Klauser, F.; Pedrozo, S. Big data from the sky: Popular perceptions of private drones in Switzerland. *Geogr. Helv.* **2017**, *72*, 231–239. [[CrossRef](#)]
62. Silvana Pedrozo and Francisco Klauser. Drones Policiers: Une Acceptabilité Controversée. 2018. Available online: <https://www.spacestemps.net/articles/drones-policiers-une-acceptabilite-controversee/> (accessed on 26 November 2018).
63. BCC. Seattle’s Ferris Wheel Hit by Drone. Available online: <https://www.bbc.co.uk/news/technology-34797182> (accessed on 26 November 2018).
64. European Aviation Safety Agency. *Advance Notice of Proposed Amendment 2015-10*; EASA: Brussels, Belgium, 2015.
65. Orwell, G. *Nineteen Eighty-Four (Also Published as 1984)*; Martin Secker and Warburg Ltd.: London, UK, 1949.
66. Lonc, B.; Cincilla, P. Cooperative ITS security framework: Standards and implementations progress in Europe. In Proceedings of the IEEE 17th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM), Coimbra, Portugal, 21–24 June 2016.
67. ETSI TC ITS. Automotive Intelligent Transport Systems. C-ITS Security. Available online: <https://www.etsi.org/technologies-clusters/technologies/automotive-intelligent-transport> (accessed on 26 November 2018).
68. PRESERVE FP7 Project. Preparing Secure Vehicle-to-X Communication Systems. Available online: <https://www.preserve-project.eu/> (accessed on 26 November 2018).
69. Petit, J.; Schaub, F.; Feiri, M.; Kargl, F. Pseudonym Schemes in Vehicular Networks: A Survey. *IEEE Commun. Surv. Tutor.* **2015**, *17*, 228–255. [[CrossRef](#)]
70. Gisdakis, S.; Laganà, M.; Giannetsos, T.; Papadimitratos, P. SEROSA: SERVICE Oriented Security Architecture for Vehicular Communications. In Proceedings of the IEEE Vehicular Networking Conference (VNC), Boston, MA, USA, 16–18 December 2013.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).