Metadata of the chapter that will be visualized in SpringerLink

Book Title	Automated Technology for Verification and Analysis
Series Title	
Chapter Title	Can We Trust Autonomous Systems? Boundaries and Risks
Copyright Year	2019
Copyright HolderName	Springer Nature Switzerland AG

Corresponding Author	Family Name	**Sifakis**
	Particle	
	Given Name	**Joseph**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Univ. Grenoble Alpes, Verimag laboratory, Bâtiment IMAG
	Address	700 avenue Centrale, 38401, St Martin d'Hères, France
	Email	joseph.sifakis@imag.fr

Abstract

Can we trust autonomous systems? This question arises urgently with the perspective of massive use of AI-enabled techniques in autonomous systems, critical systems intended to replace humans in complex organizations.

We propose a framework for tackling this question and bringing reasoned and principled answers. First, we discuss a classification of different types of knowledge according to their truthfulness and generality. We show basic differences and similarities between knowledge produced and managed by humans and computers, respectively. In particular, we discuss how differences in the system development process of knowledge affect its truthfulness.

To determine whether we can trust a system to perform a given task, we study the interplay between two main factors: (1) the degree of trustworthiness achievable by a system performing the task; and (2) the degree of criticality of the task. Simple automated systems can be trusted if their trustworthiness can match the desired degree of criticality. Nonetheless, the acceptance of autonomous systems to perform complex critical tasks will additionally depend on their ability to exhibit symbiotic behavior and allow harmonious collaboration with human operators. We discuss how objective and subjective factors determine the balance in the division of work between autonomous systems and human operators.

We conclude emphasizing that the role of autonomous systems will depend on decisions about when we can trust them and when we cannot. Making these choices wisely goes hand in hand with compliance with principles promulgated by policy-makers and regulators rooted both in ethical and technical criteria.

Keywords Autonomous systems - Knowledge - Truthfulness - Trustworthiness

Can We Trust Autonomous Systems? Boundaries and Risks

Joseph Sifakis(✉)

Univ. Grenoble Alpes, Verimag laboratory, Bâtiment IMAG, 700 avenue Centrale,
38401 St Martin d'Hères, France
joseph.sifakis@imag.fr

Abstract. Can we trust autonomous systems? This question arises urgently with the perspective of massive use of AI-enabled techniques in autonomous systems, critical systems intended to replace humans in complex organizations.

We propose a framework for tackling this question and bringing reasoned and principled answers. First, we discuss a classification of different types of knowledge according to their truthfulness and generality. We show basic differences and similarities between knowledge produced and managed by humans and computers, respectively. In particular, we discuss how differences in the system development process of knowledge affect its truthfulness.

To determine whether we can trust a system to perform a given task, we study the interplay between two main factors: (1) the degree of trustworthiness achievable by a system performing the task; and (2) the degree of criticality of the task. Simple automated systems can be trusted if their trustworthiness can match the desired degree of criticality. Nonetheless, the acceptance of autonomous systems to perform complex critical tasks will additionally depend on their ability to exhibit symbiotic behavior and allow harmonious collaboration with human operators. We discuss how objective and subjective factors determine the balance in the division of work between autonomous systems and human operators.

We conclude emphasizing that the role of autonomous systems will depend on decisions about when we can trust them and when we cannot. Making these choices wisely goes hand in hand with compliance with principles promulgated by policy-makers and regulators rooted both in ethical and technical criteria.

Keywords: Autonomous systems · Knowledge · Truthfulness · Trustworthiness

1 Introduction

Can we trust autonomous systems? This recurrent question arises quite often be-cause of their increasing importance in our everyday and future lives. Of course, we trust automated systems, as they are ubiquitous in services, devices

Y.-F. Chen et al. (Eds.): ATVA 2019, LNCS 11781, pp. 1–14, 2019.
https://doi.org/10.1007/978-3-030-31784-3_4

and appliances striving for enhanced quality of life and resource management. Nonetheless, for autonomous systems, trustworthiness becomes a major concern. They make massive use of machine learning techniques while they are highly critical as they are supposed to replace human agents in large organizations such as transport systems, smart factories, and energy production and distribution systems. Autonomous systems have already replaced to a great extent decision-making in investment markets and especially with respect to asset management (robo-advisors).

Autonomous systems significantly differ from existing automated systems in the following three key characteristics:

1. Autonomous systems deal with many different possibly conflicting goals, which is necessary for achieving adaptive behavior. This reflects the trend of transitioning from "narrow AI" or "weak AI" to "strong" or "general" AI. There is a big difference between a chess playing robot pursuing a single well-defined goal and a self-driving car that should adaptively deal with a large variety of goals including short term goals (avoiding collision and trajectory tracking) as well as longer term goals such as reaching a destination achieved by combining various intermediate maneuver goals.
2. Autonomous systems have to deal not only with a great variety of known environment configurations, but also with ones for which there is no explicit specification. This is due to the surge of cyber-physical environments: agents are sensitive to a multitude of conditions regarding the objects they need to manipulate and those that may interfere with their tasks. Another source of unpredictability is increased mobility and geographical distribution. Autonomous systems are naturally distributed which implies uncertainty on their global state and requires specific mechanisms and computational overhead to cope with it.
3. Autonomous systems are intended to accomplish complex and highly critical missions and their failure may seriously endanger their environment. It is thus desirable that in case of deviation from their normal behavior, a human operator could override their decisions and bring the system into a failsafe state. For this to be achievable, special care should be taken at design time to equip systems with adequate interaction protocols and interfaces to allow a safe transition from automated to manual regime. An alternative mode of collaboration is that the system proactively asks a human operator to take over when it diagnoses a potentially dangerous situation.

In [14] we provide an architectural characterization of the behavior of autonomous agents as the combination of five basic functions. Perception and Reflection, allow achieving situational awareness. The combination of Goal Management and Planning allows achieving adaption that depending on the perceived situation, selects relevant goals and generates corresponding action plans. A fifth function deals with the creation and handling of different types of knowledge that is essential for self-awareness and self-adaptation.

This characterization provides insight about the distinction between automated and autonomous systems. A thermostat, a lift or a flight controller are automated systems because they operate in well-defined environments that do admit a simple interpretation. They additionally pursue simple and well-defined goals and their corresponding decision process is a controller defined at design time. On the contrary, autonomous systems should exhibit self-awareness and self-adaptation for which knowledge production and management is instrumental. Although our characterization is abstract and implementation-agnostic, autonomic behavior cannot be effectively achieved without extensive use of data-based techniques and machine learning, in particular.

The use of data-based techniques in autonomous systems currently challenges our ability to provide conclusive evidence that we meet critical trustworthiness requirements. Systems engineering comes to a turning point, as traditional model-based design methodologies are not applicable to autonomous systems. Moreover, an important trend is the end-to-end development of autonomous systems based exclusively on machine learning techniques, e.g. self-driving systems providing steering angle and acceleration/deceleration from video information [3,16]. For these systems, validation is possible only by testing which cannot match the level of confidence achieved by model-based design methodologies [14].

What are the basic criteria for deciding whether a given task can be fully automated? Our analysis links truthfulness of knowledge about the behavior of a system and the resulting system trustworthiness. It comprises two steps.

The first step involves a classification of different types of knowledge according to their truthfulness and generality. We show basic differences and similarities between knowledge produced and managed by humans and machines, respectively. In particular, we note that model-based knowledge generated by algorithms can have the status of mathematical knowledge when rooted in rigorous semantics. On the contrary, data-based knowledge of neural systems is implicit empirical knowledge. It differs from scientific knowledge in that it allows prediction without understanding. We examine to what extent a principled method is applicable to machine learning techniques and highlight difficulties for achieving explainability.

The second step provides a framework allowing reasoned and comprehensive analysis of the problem whether we can trust an autonomous system for the execution of a particular task. We study the interplay between two main factors: (1) the degree of trustworthiness achievable by the system accomplishing the task; and (2) the degree of criticality of the task. In the two-dimensional space defined by these two factors, systems can be trusted when the achievable trustworthiness can match the desired degree of criticality. Otherwise, fully automated solutions are not safe enough. Nonetheless, with the advent of autonomous systems, such tasks can be jointly performed by human operators and systems, if we can achieve their harmonious and safe collaboration. We show in particular, how objective and subjective factors can influence the division of work between humans and machines.

We conclude emphasizing that the role of autonomous systems will depend on choices we make about when we trust them and when we do not. Making these choices wisely, goes hand in hand with compliance with principles rooted both in ethical and scientific criteria.

2 About Knowledge

2.1 The Truthfulness of Knowledge – A Hierarchical Classification

We consider that knowledge is truthful information which when embedded into the right network of conceptual interrelations can be used either to understand a subject matter or to solve a problem. We discuss key characteristics of knowledge and criteria that determine its truthfulness and value in use.

According to our definition, knowledge has a dual nature. It allows both situational awareness and decision-making. Thus it is crucial for perception and interpretation of the real world but also for acting on the world in order to achieve specific goals.

Knowledge can have different degrees of truthfulness and generality. It spans from factual information, to general empirical knowledge, scientific knowledge and mathematical knowledge. An important distinction is the one between empirical and non-empirical knowledge. Empirical knowledge is acquired and developed from experience. It requires thorough validation to check that it is consistent with observation and measurement. On the contrary, non-empirical knowledge is deemed independent of experience. Its truthfulness depends only on logical reasoning, while empirical knowledge is the result of (a logically arbitrary) generalization and can be falsified. It comprises in particular mathematical knowledge, theory of computing and any kind of knowledge rooted in a semantically sound framework. The Pythagorean Theorem or Gödel's theorems are "eternal" truth depending on the axioms underlying Euclidean Geometry and arithmetic, respectively.

The difference between these two types of knowledge reflects two radically different approaches for its production. One is a purely logical construction while the other concerns information extracted from observations and experimental data. Figure 1 proposes a classification allowing a comparison between types of knowledge produced and managed by machines and humans.

The most common kind of empirical knowledge is explicit knowledge about facts characterizing situations of the world at a certain time and place e.g. "the temperature in Paris today is 24 °C" or "the battle of Waterloo took place on Sunday, 18 June 1815". Factual knowledge is of limited generality but indispensable for situational awareness.

General empirical knowledge is the result of generalization and abstraction of factual knowledge. It comprises in particular, implicit empirical knowledge which involves learning and skills but not in a way that can be explained and analyzed. This is the most common knowledge humans use to walk, speak, play instruments, dance, etc. It is produced and managed by automated (non-conscious) effortless fast thinking (System 1 of thinking according to D. Kahneman's terminology [9]). When we walk, our mind solves a very hard computational problem whose explicit modeling would involve dynamic equations describing the kinetics of our bodies. Note that neural systems produce and handle implicit empirical knowledge. They learn to distinguish "cats from dogs" exactly as kids do. This type of knowledge also comprises statistical knowledge and knowledge produced using data analytics techniques.

Scientific and Technical knowledge is past empirical knowledge that has been processed and systematized through the use of models. Scientific knowledge allows understanding the physical world while technical knowledge allows building new products or processes based on scientific knowledge e.g. engineering constructions. The big difference between implicit knowledge and scientific and technical knowledge is that the letter is model-based and thus it is amenable to falsification analysis, which drastically improves confidence in its truthfulness.

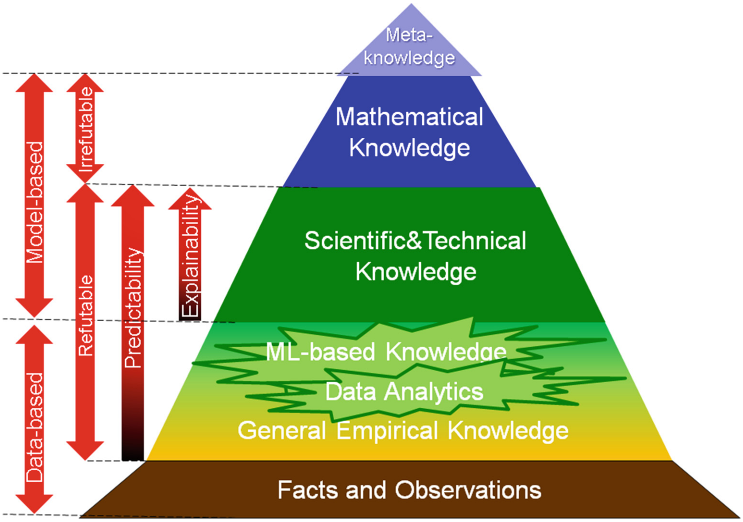


Fig. 1. The knowledge pyramid

As explained, non-empirical knowledge is model-based knowledge rooted in logical rules.

Finally, meta-knowledge is knowledge about how to deal with knowledge. It allows combining various kinds of knowledge for situational awareness and decision-making. It includes design methodologies, problem-solving techniques, data acquisition and analysis techniques. It also includes non-formalized knowledge related to various jobs and skills.

Note that mathematical knowledge as well as scientific and technical knowledge are model-based. Humans produce this type of knowledge by slow conscious, effort-ful procedural thinking (System 2 of thinking according to D. Kahneman's terminology [9]). Conventional computers can handle this type of knowledge, when it is adequately formalized, and produce new knowledge e.g. by executing algorithms. There is a remarkable similarity between the two types of thinking (fast and slow thinking) and the two types of computing (conventional algorithmic and neural computing). Both slow procedural thinking and ordinary computing are model-based in the sense that it is possible to produce a model explicating step-by-step the underlying computational process. On the contrary, both neural computing and fast thinking emerge as the result of some learning process that does not rely on any explicit procedural model.

2.2 Scientific vs. Machine-Learning Knowledge

We discuss differences in the production processes of scientific and machine learning knowledge respectively, and how these affect their corresponding degree of truthfulness.

The scientific method consists in developing knowledge that faithfully accounts for experimental observations. Scientific discovery is the result of the implicit learning mental process of an experimenter who builds a model allowing predictability and explainability.

Neural systems generate knowledge as the result of a long training process with experiential data. Their explainability is the object of active research investigating various definitions of the concept and associated explanation techniques, e.g. [6, 11]. For our comparison we simply consider that explainability implies the existence of an analyzable model matching the observed behavior.

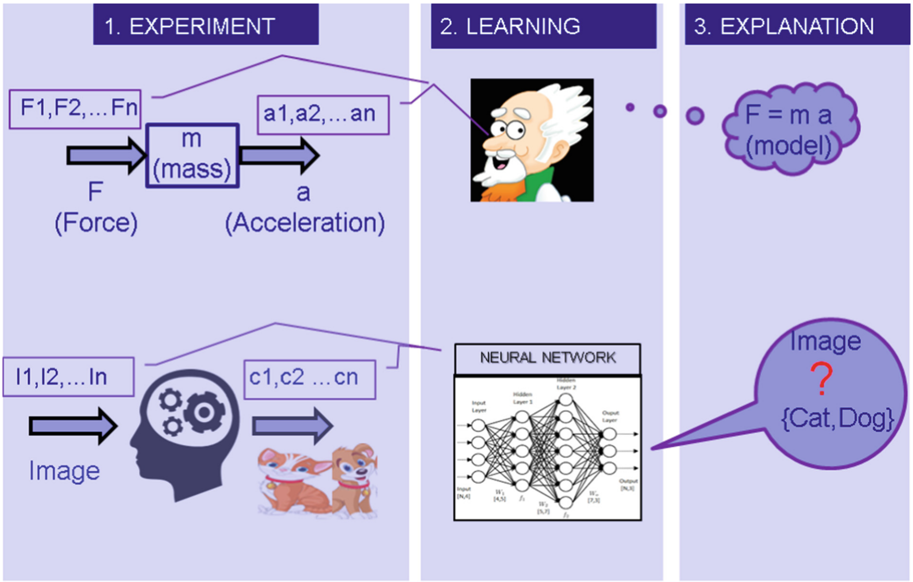


Fig. 2. Comparing scientific and machine learning based knowledge

Figure 2 illustrates a comparison between the scientific approach for studying a physical process (mass acceleration a by force F) and the technical approach for learning a mental process (distinguishing between images of cats and dogs). They both have a common purpose to characterize the input/output behavior of the considered processes and achieve predictability: they are able to guess the response of the process for a given input.

Scientific knowledge is empirical knowledge represented by analyzable models whose behavior can be studied and tested. In that manner, observations and experimental data take a higher value and generality. Scientific discovery is not possible without adequate models. We know that Newton has developed infinitesimal calculus so that he could formulate his laws. Our difficulty with fully understanding and predicting complex phenomena such as social, meteorological and economic, does not necessarily imply that they do not follow laws, but simply that we do not have the adequate models explaining of the observed data. Additionally, the development of scientific knowledge requires that the models can be analyzed to study their behavior and extract significant properties. So it implies some computational complexity that may limit explainability.

Similarly, the machine-learning paradigm involves an experimental step followed by a learning step applied to a neural network. It consists in adjusting weights of the network so that it computes a function fitting as closely as possible the observed behavior. The so obtained neural system allows predictability with a probability depending on the degree of training. The application of the third step to find a model explaining the network behavior, is a largely open problem, in particular for neural systems that emulate mental processes dealing with hard to formalize concepts of the natural language. Is there a rigorous model relating images to cats and dogs? Nonetheless, explainability seems feasible for specific classes of neural networks dealing with physical entities for which it is possible to characterize rigorously their I/O behavior e.g. by sets of constraints as for example in [10]).

3 Trustworthiness vs. Criticality

3.1 System Trustworthiness

Our trust in systems depends on the truthfulness of our knowledge about their components and the way they are built.

Trustworthiness characterizes system resilience to any kind of hazard including [12,15]: (a) software design and implementation errors; (b) failures of the execution infra-structure and system peripherals; (c) interaction with potential users including erroneous actions and threats; and (d) interaction with the physical environment including disturbances.

Note that trustworthiness concerns not only functional properties but also general non-functional properties including safety and security. It characterizes the whole system's computing environment. Among the possible hazards, only software design errors and defects require functional validation. The others require the analysis of a system model in interaction with its physical and human environment. Trustworthiness depends on both technical and subjective factors. Technical trustworthiness assessment is a complex task involving separate evaluation of functional correctness for a nominal behavior against a set of requirements [15]. It is followed by a risk analysis of potential issues that could affect the system safety and security. Trustworthiness is especially characterized by the probability that events with catastrophic con-sequences occur – as an example, this probability for transport category aircraft should be less than 10^{-9} failures per flight hour. The assessment of system trustworthiness involves three different levels of knowledge.

1. *Irrefutable evidence* that a mathematical system model meets given requirements. This is the type of knowledge obtained by analysis of system models. In that manner, we can estimate the energy consumption of a circuit model, compute a program in-variant, or show that the RTL model of a piece of hardware computes a given function.
2. *Conclusive evidence* that a system meets given requirements. This is the type of knowledge obtained as the result of a two-step process. The first step involves the construction of a mathematical model of the system and checking that the model is faithful, i.e. each true statement about the model holds for the real system. Then, the so obtained model is analyzed to get irrefutable evidence that will hold for the real system under the assumption of model faithfulness. Conclusive evidence is the most truthful knowledge one can get about real systems. It is often required by critical systems standards that explicitly recommend the use of model-based design techniques.
3. *Sufficient evidence* that a system implementation passes a test campaign. Testing allows discovering defects but cannot guarantee absence of defects, which is possible by conclusive evidence. Its efficiency can vary depending on the rigorousness of test coverage criteria. This type of experimental validation suffices only for non-critical systems.

Lack of explainability of neural systems implies that our knowledge about their behavior is restricted to sufficient evidence. On the contrary, for systems developed according to rigorous model-based approaches, the three types of knowledge are equally useful to ascertain their trustworthiness. Reasoning on system models allows irrefutable guarantees and strong predictability. Verification of system components and of the system development process brings conclusive evidence about correctness with respect to requirements. Finally system testing plays a complementary role. It brings additional sufficient evidence about the actual system implementation by exercising the code generated by a compiler from the application software and running in a given execution environment.

3.2 The Automation Frontier

How we decide whether a system can be trusted to perform a given task? Our decision depends on two main factors: (1) System trustworthiness; (2) Task criticality, which characterizes the severity of the impact of a failure in the fulfilment of the task.

We assume that system trustworthiness varies in the interval $[0, 1]$. The highest trustworthiness corresponds to systems that in all cases would behave as expected while the lowest to systems that exhibit completely random behavior.

Task criticality characterizes pure functionality provided by the system and is completely independent from implementation issues. Driving a car, operating on a patient, and nuclear plant control involve intrinsic risks that do not depend on the way these tasks are carried out and the means employed.

We similarly assume that the degree of task criticality is in the interval $[0, 1]$. The highest criticality corresponds to catastrophic errors with costly consequences. The lowest criticality means indifference to errors in the performance of tasks. Further-more, we assume that there is a monotonic correspondence between the achievable system trustworthiness and the required task criticality: a given trustworthiness level allows satisfaction of a corresponding criticality requirement.

Consider the two-dimensional space defined by the two quantities, system trust-worthiness and task criticality. A system for a given task is represented as a point in this space (Fig. 3).

Based on these definitions, the answer to the problem is simple: if the trustworthiness of a system realizing a given task is greater than the required degree of criticality, then the system can be trusted. Otherwise, it is not reasonable to trust the system; the task may be assigned to skilled human operators or cannot be performed by either humans or systems.

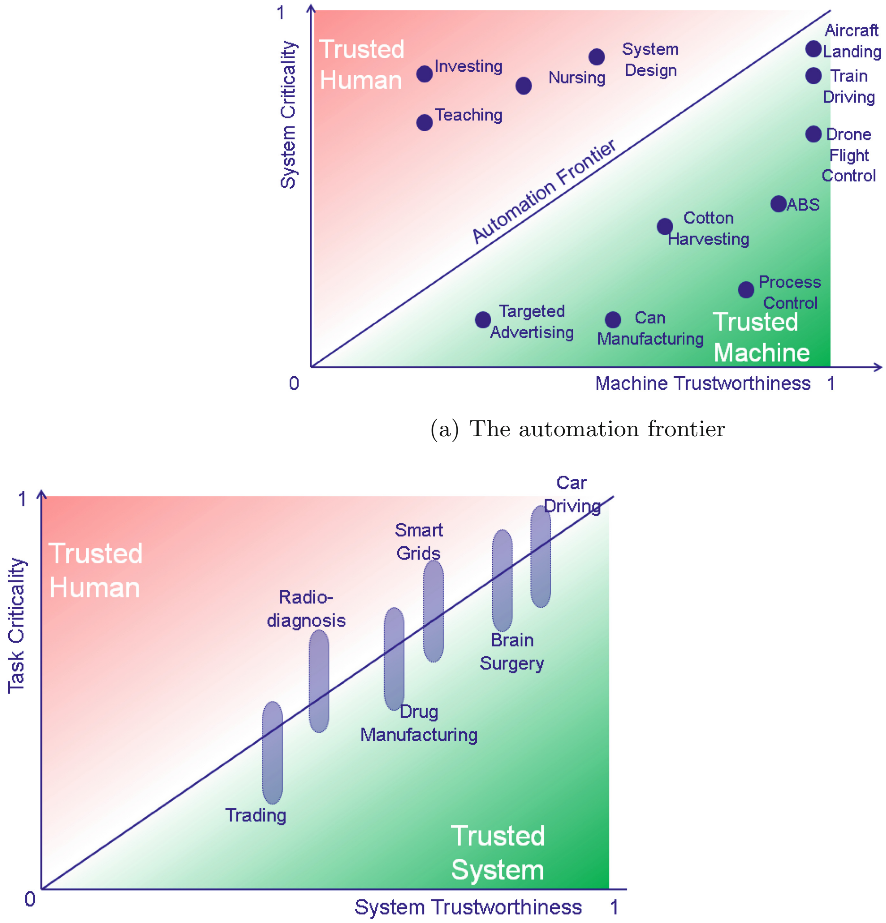
Figure 3(a) shows automated systems that are trusted because they meet this requirement. It also shows tasks assigned to humans and for which the achievable trustworthiness cannot match the required criticality level e.g. teaching. For the sake of simplicity, we assume that each trustworthiness level matches exactly the same level of criticality. In that case, the angle bisector defines the frontier of automation, which separates the space in two regions: one where machines can be trusted (below the frontier) and another where humans may be more trusted than systems for the same task. With increasing automation human-operated tasks cross the automation frontier and pass from the red to the green region.

As explained, the advent of autonomous systems will allow the automation of complex tasks that are currently entrusted to humans in large organizations. The transition to fully autonomous systems will be progressive and will be eventually completed in some distant future. Meanwhile, the challenge is how to achieve symbiotic autonomy [4] by determining the appropriate balance in the division of work machines between humans and computers. This idea illustrated in Fig. 3(b) is reflected in the concept of degree of autonomy e.g. SAE definitions for self-driving cars [1].

3.3 Other Factors Shaping the Automation Frontier

It is important to note that the defined ideal automation frontier can be distorted by other objective or subjective factors (Fig. 4).

One factor is the big difference in performance between machines and humans. If the task performed by the system is not critical, we may use auto-



(a) The automation frontier

(b) Degree of automation for autonomous systems

Fig. 3. The automation frontier and the degree of automation (Color figure online)

mated systems above the automation frontier because the gains in performance can be substantial and compensate a relatively high failure rate. Today, we use many automated services such as internet bots that perform repetitive non critical tasks at a much higher rate than would be possible for humans.

On the contrary, for high criticality levels, people have the tendency to accept and excuse human mistakes if they understand the circumstances which shaped the wrongful or reckless behavior.

As a rule, the public opinion is more unforgiving for system failures than for human errors e.g. accidents caused by self-driving car vs. accidents caused by human drivers. This bias in favor of humans shapes the automation frontier in

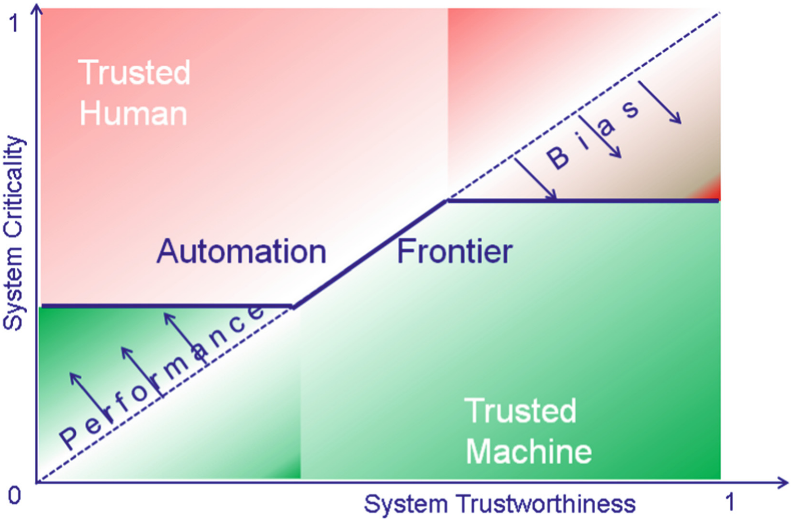


Fig. 4. Shaping factors of the automation frontier

the opposite manner. Even if systems may be as trustworthy as humans, their acceptance to perform highly critical tasks would always be questioned.

We have assumed that system trustworthiness and criticality are quantities rigorously and indisputably defined. We explained that conclusive evidence is achievable only when specific model-based system design methodologies are applied. Such methodologies fall short for autonomous systems due to several reasons including the non-predictability of their environment and of course, the inevitable use of data-based techniques.

Besides these considerations, it should be emphasized that trustworthiness has a subjective and social dimension. We should not ignore the role of institutions that directly or indirectly contribute to shaping public perceptions about what is true, right, safe, etc. in modern societies. It is not enough to build a system complying with the acknowledged rules of the state of the art. Care should be taken that the compliance of the construction process can be checked by independent experts [5]. Certification of critical systems should remain the prerogative of independent agencies according to well-founded standards requiring conclusive model-based evidence.

4 Discussion

We presented a classification of knowledge depending on its truthfulness and resulting types of evidence about system trustworthiness. This classification should not be associated with a judgment of value. Scientific knowledge is deemed more truthful than general empirical knowledge, but its development and application are limited to formal domains of discourse. Machine learning techniques are indispensable for autonomous systems because they can effectively deal with concepts of natural languages. Furthermore, we have argued that model-based knowledge does not suffice to ascertain system trustworthiness. Reasoned development of empirical knowledge is also important to check the consistency of implementations and links to the physical world.

The challenge for autonomous systems is to consistently combine these different types of knowledge and bridge the gap by cross-fertilization of approaches. How to enhance truthfulness of empirical knowledge based on rigorous qualitative or quantitative criteria? Clearly, monolithic end-to-end solutions are not amenable to analysis and testing remains the only way to assess trustworthiness. Using modularity principles as recommended by standards such as ISO 26262, allows mastering design complexity by restricting the size of components and maximizing the cohesion within a component [7]. Modular architectures could involve both data-based and model-based components seeking trade-offs between trustworthiness and performance.

Furthermore, data-based techniques can be profitably used to overcome limitations of the scientific method. One well-identified limitation comes from the cognitive complexity of the relations that our mind can apprehend: we can deal with relations of rank up to five (one predicate + four arguments) [8]. This is especially reflected in the fact that scientific theories involve a limited number of fundamental independent concepts. Einstein was considering that we are lucky that basic physical laws are simple enough ("The most incomprehensible thing about the universe is that it is comprehensible."). The impressive success of physical sciences often makes us believe that everything should be explainable in terms of formal theory, simple enough to be developed by humans. However, our current lack of holistic understanding of complex phenomena does not necessarily mean that they are not subject to laws. They may well obey to laws that we cannot discover as their complexity exceeds our cognitive capabilities. There are many examples where computers contribute to the analysis and deeper understanding of complex phenomena via the combined use of data analytics and learning techniques e.g. [13]. The challenge is to defeat cognitive complexity by achieving computer-assisted development of scientific knowledge.

The role of autonomous systems will depend on our decisions about when we trust them and when we do not. Making these choices wisely depends on two factors. The first factor is our ability to appreciate, based on well-founded criteria, whether and to what extent we can trust knowledge produced by computers. Our analysis emphasizes the importance of two inter-related concepts: truthfulness of knowledge about how a system behaves and the resulting system trustworthiness. We need new theoretical foundation and technology for the evaluation of trustworthiness of autonomous systems that integrate both model-based and data-based components. Such results could be a basis for the definition of standards for the development and use of autonomous systems (as we do for all artifacts from toasters to bridges and aircraft). The current trend for self-regulation and self-certification should be considered as a temporary stopgap rather than the definitive answer to the trustworthiness issue.

The second factor is increased social awareness and sense of political responsibility. It would be good to apply the precautionary principle that already underlies laws and regulations in European Union: when computers are part of critical-decision processes we should make sure that their judgment is safe and fair [2]. This principle should be embodied in laws and regulations governing their development and deployment.

We are on the verge of a great knowledge revolution. We should be vigilant and question the use of machine-produced knowledge allowing predictability without under-standing critical decision processes. I believe that the threat is not that computer intelligence surpasses human intelligence and that computers take power and control over human societies by hatching a plot. The real danger comes from the massive re-placement of accountable and responsible human operators in critical decision processes. Let us hope that we will not grant the power of decision to autonomous systems without rigorous and strictly grounded guarantees under the pressure of economic interests and on the grounds of an ill-understood performance benefit.

References

1. National Highway Traffic Safety Administration, et al.: Federal automated vehicles policy: accelerating the next revolution in roadway safety. US Department of Transportation (2016)
2. Benkler, Y.: Don't let industry write the rules for AI. Nature **569**(7755), 161–161 (2019)
3. Bojarski, M., et al.: Explaining how a deep neural network trained with end-to-end learning steers a car. arXiv preprint arXiv:1704.07911 (2017)
4. Dambrot, S.M., de Kerchove, D., Flammini, F., Kinsner, W., Glenn, L.M., Saracco, R.: IEEE symbiotic autonomous systems white paper II (2018)
5. De Millo, R.A., Lipton, R.J., Perlis, A.J.: Social processes and proofs of theorems and programs. Commun. ACM **22**(5), 271–280 (1979)
6. Doran, D., Schulz, S., Besold, T.R.: What does explainable AI really mean? A new conceptualization of perspectives. arXiv preprint arXiv:1710.00794 (2017)
7. Frtunikj, J., Fürst, S.: Engineering safe machine learning for automated driving systems. In: Proceedings of the 2019 Safety-Critical Systems Symposium, pp. 115–133 (2019)
8. Halford, G.S., Baker, R., McCredden, J.E., Bain, J.D.: How many variables can humans process? Psychol. Sci. **16**(1), 70–76 (2005)
9. Kahneman, D.: Thinking, Fast and Slow. Macmillan, London (2011)
10. Katz, G., Barrett, C., Dill, D.L., Julian, K., Kochenderfer, M.J.: Reluplex: an efficient SMT solver for verifying deep neural networks. In: Majumdar, R., Kunčak, V. (eds.) CAV 2017. LNCS, vol. 10426, pp. 97–117. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-63387-9_5
11. Lipton, Z.C.: The mythos of model interpretability. arXiv preprint arXiv:1606.03490 (2016)
12. Neumann, P.G.: Trustworthiness and truthfulness are essential. Commun. ACM **60**(6), 26–28 (2017)
13. Rouet-Leduc, B., Hulbert, C., Lubbers, N., Barros, K., Humphreys, C.J., Johnson, P.A.: Machine learning predicts laboratory earthquakes. Geophys. Res. Lett. **44**(18), 9276–9282 (2017)

14. Sifakis, J.: Autonomous systems-an architectural characterization. arXiv preprint arXiv:1811.10277 (2018)
15. Sifakis, J., et al.: Rigorous system design. Found. Trends® Electron. Des. Autom. **6**(4), 293–362 (2013)
16. Zeng, W., et al.: End-to-end interpretable neural motion planner. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8660–8669 (2019)