A publication of ISCA*:
International Society for Computers
and Their Applications

INTERNATIONAL JOURNAL OF COMPUTERS AND THEIR APPLICATIONS

TABLE OF CONTENTS

*** "International Journal of Computers and Their Applications is abstracted and indexed in INSPEC and Scopus."**

Volume 24, No.4, Dec. 2017 **ISSN 1076-5204**

International Journal of Computers and Their Applications

A publication of the International Society for Computers and Their Applications

ISCA Headquarters···•···P. O. Box 1124, Winona, MN 55987 USA···•···Phone: (507) 458-4517
E-mail: isca@ipass.net • URL: http://www.isca-hq.org

Guest Editorial: On Humans, and Systems They Create

Maximilian M. Etschmaier, Kendall E. Nygard, Eugénio Oliveira,

An essential characteristic of humans, as of all living creatures, is that they transform elements of their environment, either to sustain their existence or to make it more palatable. The creator as well as the target of all transformations, therefore, is the human. Whatever may affect the transformation is an artifact created by humans. This artifact, together with the human and the elements of the environment that are being transformed, form a system. As long as such systems were relatively simple, humans were able to follow natural instincts to define and implement the transformations. It made little difference whether or not they understood that the human was included in any system.

As the nature of the transformations became more complex, targeted exploration and design efforts were needed to identify transformations that would enhance the human existence. The skills to create systems have increased over time, and so have the capabilities and complexity of systems. Human-created artifacts can now handle tasks that previously were reserved for humans. These human-created artifacts are no longer just mechanical devices that execute prescribed tasks. Increasingly, they are capable of logical reasoning and independent decision-making. They are able to recognize and interpret complex situations, even situations that could not have been anticipated by their designer, to learn from those situations, and to modify their own behavior. Thus, they are no longer mere tools of humans, or "machines" that are operated by humans, but they are integrated with humans into symbiotic systems. Only if the human is considered as a part of the system from the very beginning of the design process, is it possible to fully consider the potential of the symbiotic relationship between the human and the human-created artifact; and to arrive at a system that is internally consistent, and potentially most effective and efficient.

A design paradigm suitable for designing such systems will be radically different from what served us well in the past: in addition to the artifacts that were created before and that impact a design, the designer will need to understand the nature of the human in the system, including the human's interest, capabilities, strengths and weaknesses, as well as motives, preferences and expectations. A prerequisite is that designers ultimately understand their own selves. This gives new meaning to qualities like introspection, imagination, and creativity. It is that which permits the designer to identify some initial purpose and boundary for a system and iteratively evolve it until it reaches a state of balance where the purpose is enclosed by the boundary to the extent that is consistent with the relationship of the system with other systems.

It also means that system functions and events that previously were seen in their mechanical (or physical) dimension alone, need to also be viewed from a human perspective. From a human perspective, the occurrence of critical system failures (i.e., failures that have unacceptable consequences, such as a loss of life) must not be tolerated. As the human ultimately is in control of a system through the design as well as through operational decisions, any critical failure can no longer be trivialized as an accident, but should be recognized as a loss of control of the system by a human.

There are other, more dangerous ways in which the human can lose control of a system. If the roles (functions) of the different system components are not assigned in accordance with the capabilities of the components, the system is not balanced internally. Some functions may not receive sufficient weight in the process of internal decision-making, while others may be given excessive power. As a result, the system might be moving in a direction that ultimately may imperil its existence. In particular, if the imbalance occurs at the boundary between human and "mechanical" elements, the human side may not receive the information or possess the decision-making power to keep up with processes that occur within the "mechanical" side and the human may, therefore, lose the ability to effectively control it. This can happen without the "mechanical" element rebelling against the human as envisioned by some researchers in artificial intelligence.

Another way in which the human can lose control over a system is if the system is not in balance with the environment. For example, the system may possess the power to destabilize the environment, but lack the ability to reestablish a viable equilibrium. If the environment that is affected extends over the globe, as is the case in the current climate crisis, humanity may lose the basis for its very existence.

This then appears to be what is required of any system design:

- That the system is in balance with the environment
- That the system is balanced internally
- That the system structure inherently eliminates the possibility of the human losing control to the "machine."

As the human is more tightly integrated into the system, and the human-created artifacts consist more and more of logical constructs, systems discussed so far are approximating a class of systems that include different types of human-created artifacts: constructs of human thought that define human behavior, curiosity and creativity, ethical principles, and other norms of social behavior. Although these systems are the domain of different professional and scientific disciplines, the methodologies for the analysis and design will converge, especially as the awareness increases that these other artifacts also require a conscious effort in design in order to be consistent and effective. Similarly, the qualifications required of the designers in any domain will converge to one design paradigm. As engineers we might view this as another engineering discipline. Sociologists may differ on this, but, whatever the name and the academic home, it will be important that there is a cross-fertilization between systems design conceptualization and execution across all domains.

Situations where the human lost control of a system they created, and are part of, have occurred throughout history. Natural environments have been destroyed, depriving humans of their livelihood. Social and political systems, including large empires, have collapsed because of ill-conceived human-created artifacts like organizational structures and rules and standards of ethics, commerce, and other forms of human intercourse. In the past, humans have been able to escape the most-dire consequences of losing control by moving elsewhere. Failed technological systems were abandoned. Failed social and political systems were replaced by new ones. Today, the space for human action is increasingly constrained by the finiteness of the planet. As systems increasingly are spanning the entire planet, the consequences of loss of human control over human-created systems leave the human without alternatives. The loss of control thus is irrevocable. And it is that which makes the study of human-machine symbiotic systems and of humans losing control over such systems so urgent and important.

However, it is not enough to study and analyze the nature and behavior of such systems. Since the threat to human control over systems derives from human-created artifacts, it is necessary to redesign or replace those artifacts with new ones. The focus of this endeavor on the design of a system reflects the interest of an engineer. It is different from the interest of a scientist, who extracts essential information about the nature of a system, its behavior, as well as the properties of the system components. However, both are creating models of systems, with the models of the scientist providing the foundations for the designs of the engineer.

This special issue of the Journal combines both perspectives. It reflects the results of discussions within the International Society for Computers and their Applications (ISCA) over several years. It is hoped that it provides a foundation for future discussions within the society and within all relevant scientific and professional associations.

Can Humans Stay in Control of Systems They Create?

Maximilian M. Etschmaier
San Diego State University, San Diego, CA USA

Abstract

The nature of the relationship between humans and artifacts they create ("machines") is examined. It is shown that the process of machines increasingly taking on roles that were previously reserved for humans is an essential part of the evolution of humanity. However, it is considered unlikely that, by virtue of massive processing power and sophisticated artificial intelligence algorithms, machines will ever assume a life of their own. Instead, as algorithms become too complex for humans to comprehend, serious risks arise from the fact that machines will do exactly what humans tell them to do instead of what they mean them to do.

A much bigger threat to humanity is shown to emerge from machines capable of influencing the human mind. Such machines have long existed in the context of marketing and political campaigns. It is suggested that the power they gained through global real-time access to information and communication, especially in the form of social media, enables them, like parasites, to seize societies (or humanity as a whole) and with them form a new form of life. This would be the ultimate way in which humanity can lose control. And it would be irreversible.

Some strategies are identified that may minimize the risk of humans losing control in either scenario.

Key Words: Human-machine systems, artificial intelligence, robots, social media, humans losing control, machines controlling minds, purposeful systems.

1 Introduction

The 1970's and 1980's saw intensive discussion of an impending revolution through artificial intelligence. Computers would be built with "intelligence" that would exceed that of the human brain. The sheer mass of intelligence would eventually provide these computers with the capability of autonomous reasoning – a life of its own. That would put them in a position to control their own actions, to grow by themselves, and to create copies of themselves. As a result, they would at least challenge humans for dominance. As there would not be any limit to their growth, victory would be assured and humans would be forever dominated by their own creation [1].

Things did not quite turn out as expected. A proof that the concentration of a large amount of power to process logical operations ("intelligence") would, by itself, lead to the emergence of independent systems with life-like properties has not emerged. And it is considered ever less likely that such systems will emerge. However, the capacity of computers and the power of computer programs have increased to the point where computers can, by themselves, perform tasks that previously could only be done by humans. And systems are increasingly emerging that, while not having a life of their own, can act autonomously to follow a human-provided program. Robots are being developed for many situations from manufacturing to transportation and to warfare.

It might be expected that such robots cannot ever be a threat to humans. After all, they do exactly what they are told to do. However, as Wiener [14] has pointed out, that is exactly the problem. What humans tell them to do, may not be what they mean them to do. And if these robots are autonomous, there is no way to stop them. Civilian robots may cause horrible accidents, and killing robots may turn on friendly troops. Increasingly, leaders in civilian and military applications are urgently warning of the dangers [2-3, 7, 10]. If it is not the accumulation of intelligence itself that causes humans to lose control of systems they create, then it must be features of the design and the design process that lets control to silently slip out of humans' hands. And by the time humans realize that they have lost control, systems may have become so intractable that it may be too late to regain control.

The most recent developments in mathematics, and information and communication technology are not only leading to new social systems related to the organization of work and processes of production and distribution, but also make possible the formation of entirely new types of social systems around common beliefs, passions, and tastes. These social systems are entirely separate from the spatial, and to some extent, the time dimension. They integrate humans with human-created physical and logical artifacts and with the environment. The artifacts in them, like parasites, draw on the life of the human to acquire the energy of life for themselves and with it attributes of living creatures like consciousness and emotions. Such systems overcome the limitations of conventional systems of artificial intelligence and may well represent a realistic path to what the original authors of artificial intelligence were aiming for. To create or

analyze such systems, in addition to technology, it is necessary to explore the nature of humans and the processes through which they are evolving over time.

2 On Human Nature

A central feature of humans is that they consciously develop artifacts that improve their condition, both physically and socially. These artifacts change the nature of humans, thereby driving the process of their own evolution as social beings. To examine this, it is necessary to trace humanity back to its very beginning and identify the properties that started this process. We shall see that the same properties that have driven the process in the past are responsible for continuing the process, possibly to the point where humans may lose control.

It is not possible for humans to know the facts of their origin. However, there are various narratives of it that are written by humans that have been accepted, each by a different human subgroup. These narratives, even as many are presumed to be divine-inspired, are human-created artifacts as any work of history. By virtue of their acceptance, they become objective facts [4, 11]. And as that, they define humans as conscious beings, who reflect on their actions, experiences and possess a drive to explore and conquer the world, and a free will that includes the freedom to turn against the creator as well as against fellow humans. But there are different creation narratives, created by different humans (prophets, thinkers, artists, scientists, …), and each accepted by a different subgroup of humanity. Each narrative, in its own way, defines a subgroup of humanity and provides the basis for the evolution of the subgroup. The subgroup and its creation narrative become inseparable.

Over time, new artifacts are added to the creation narrative and accepted. For example, language (including art and mathematics) defines thought and interactions between humans. Ethical norms define human behavior. Eventually, ideologies emerge that define how value is determined and that set rules for commerce, for the operation of political entities, and for economies. All these artifacts become one with humans. They form social entities that include and are above humans and define future developments. Each subgroup of humanity, thus, is defined by the individuals included in it and the human-created artifacts embraced by it. Each individual shared the artifacts of its group and participates in its evolution. However, different experiences and the free will may move individuals in different directions.

Two developments have greatly increased the power to get humans to create and accept new artifacts: The emergence in the late 19th century of psychology produced insight into the workings of the human mind. Although originally meant as a tool to recognize and cure mental illness it was quickly adapted to learn how to manipulate the human mind, especially in marketing of goods and services and of political ideas. The second development, in the middle of the 20th century concerned the rapid development of technologies for information processing and communication as well as techniques for automatically extracting information from large sets of data in particular the internet, social media, and data analytics.

Both developments have also made it easier to deliver artifacts of thoughts to receptive recipients, influence their thinking and behavior, and to supersede personal reasoning and commit individuals and groups to new ethical norms. This intensifies the question about truth and the objectivity of facts.

The combination of these two artifacts makes it possible for some entity (a small number of individuals) to target the minds of groups of humans (or all humanity) and, like the pied piper, lead them into total submission (i.e., to their peril). Examples are marketing and political campaigns including indoctrination.

To perpetuate control, it may be necessary to constantly update the control mechanisms. Eventually the mechanism may become so complex that the controlling entity can no longer understand it. Updates will be limited to peripheral features. The mechanism will continue to follow the program included in it and will adapt itself to the changing environment as its program prescribes. The more changes will have been executed, the less likely the controlling entity will be able to regain control. Thus, the mechanism will come to act as an autonomous agent that will possess total control over itself and lead the group controlled by it in a direction that only it decides upon. Although this may make it appear like a living creature, it remains a deterministic mechanism that follows its instructions and, for example, lacks emotions. Therein lies the danger. Especially if the mechanism is designed to play competitive games it may recognize attempts to modify its course as attacks on it and turn against any human it may perceive as a threat, and since it lacks emotions, it will not stop until it has defeated all.

Social media provides a way to imbed algorithms into a social media network. The algorithm will identify individuals that are potentially susceptible to similar ideas. It increasingly links them and at the same time reduces links to others that are not susceptible. This creates a “filter bubble,” an echo chamber that confirms the idea with the members and creates a community around the idea. As the communication among the members continues the identification with the idea intensifies. The idea and the algorithm then, once implanted, acts like a parasite that takes the energy it needs to sustain itself from its host, gradually taking control of the host. The individuals become trapped in a bubble they themselves helped create and are helping perpetuate. The algorithm, or someone controlling it, can move the bubble in directions that serve its purpose, even if that is opposed to the purpose of the people in it. The people in the bubble will not understand the difference and actively follow the purpose of the algorithm. Through the algorithm the bubble might also be set up to engage in hostility toward other bubbles orchestrating conflicts that may establish dominance, but be harmful to all humans involved.

3 The Status Quo

Today, the understanding of human psychology and computer and information technology is growing at an ever-increasing speed. Software systems, hardware embedded or not, are now capable of programing themselves in such a way that they adapt to new situations and reach decisions that could not be anticipated before. Moreover, many of these so-called intelligent systems feed on data we sometimes unconsciously are producing all the time.

Information technology provides wireless access to an ever-increasing range of public and private spaces and enables increasingly universal monitoring of private activities. An increasing range of delivery mechanisms through various information networks make it possible for individuals to spontaneously and informally communicate with each other. Remote control of devices that define the everyday life experience of individuals make decisions for them. They drive automated home appliances, present suggestions for entertainment offerings like movies, TV shows, books, and other leisure activities, shopping and the consumption of goods and services, as well as for social interaction.

The public is being seduced into providing access to most private information by piggybacking onto desirable services like intelligent home appliances, toys, private conversations via media, location services for route planning and emergency response, automatic toll collection among many others. As part of safety and security screening to catch "bad guys," they readily accept the collection of samples of finger prints, face recognition and other biometric data, and license plate monitoring. As a result, in some specific contexts, each one of us could be targeted, advised, compelled, (who knows, forced), to act, to buy, to vote as indicated otherwise.

The increasing understanding of psychology and computer and information technology has also brought about large-scale displacement of human labor across the economy, for example in manufacturing, and mining, but also in retail, engineering design, and computer programming. By enabling monopolized access to networks it has increased the concentration of providers of many kind of services. Businesses are increasingly concentrated into one-stop service centers for shopping, entertainment, tourism and transportation. The combination of psychology and technology can make changes appear palatable, even advantageous, even if they only serve the interests of a select few.

In politics, we observe a rapidly increasing effectiveness of controlling (and changing) opinions. The outcome of elections is more often the result of technology available to a campaign. The frequency of small margins of victory in strategic precincts may attest to the effectiveness and efficiency with which technology is being deployed. We also observe how readily selected artifacts that govern human behavior, e.g., ethical norms, or the value of human life can be changed as is evidenced by the increasing number of individuals who are ready to sacrifice their lives for a cause they have been convinced was noble.

Humans have created a machine (robot, algorithm) that is ready to subjugate them. Unless we take decisive action, the machine will succeed. And there will be no return. We observe how the space for productive human activity is steadily shrinking being reduced to vision, creativity, imagination, conscious reflection, and leadership in charting the future. Only humans who possess qualities to excel in these areas are assured of being able to make valuable contributions to society. The rest of the population may find themselves without any opportunity. Emerging from this could be huge social problems that may defy an equitable solution because this population, in the prevailing perspective, is literally surplus in the productive economy and society. The inescapable consequences are widening wealth disparities increasing enslavement of lower classes by an ever-shrinking class of super-rich.

Only a blind cynic could view this as a sustainable state. Humanity will continue to evolve with continued development of artifacts, groups with cohesion will grow larger, but more importantly, the bonds within them will grow ever stronger. The conflict between groups will increase. The future of humanity is at stake. Something has to be done to change things. We have tried to show that this problem was accelerated by the emergence of psychology and technology. But the effect of science and technology is only to accelerate a process that represents the essence of human existence. It appears to be human destiny to get ever closer to the abyss. It is our responsibility to chart a course that keeps us from falling off the edge.

4 The Way Forward

Two forces are responsible for driving development of the human condition: science and engineering. While there is great confusion in the use of these terms, we shall use them here in definitions that express the essential difference between them permitting us to demonstrate their interdependence and the external factors that influence them. This will permit us to identify forces that may be responsible for the direction of the human enterprise.

We define "science" as the domain of learning that is concerned with the creation of new knowledge, and "engineering" as the domain that uses existing knowledge to create new systems that serve a recognizable purpose. This definition does not limit the objects of engineering and science to the physical domain, but equally includes logical domains as well as the social, economic and political domains.

In principle, science is driven by essential human desires for knowledge (curiosity), and by the recognition of opportunities that particular knowledge could provide for the improvement of the human condition. Engineering is driven by the recognition of benefits of new and improved artifacts (systems) that may arise from the knowledge discovered by science, as well as of artifacts (systems) that could be developed if certain new knowledge could become available

from science. Thus, science and engineering are in a relationship of mutual interdependence and, together, chart the path into the future.

However, both are controlled by humans who are influenced by economic and social rewards. And through that, they are influenced by the forces of the environment they are embedded in. These result from the constellation of political, economic, academic, ideological, and other powers and institutions. The desire for "success" (however defined) will influence the selection of projects to work on. It will also cause them to join in groups of various kinds to exert power in their environment to create specialty areas of knowledge with their own language within which they can control the criteria of success. The power of these areas of knowledge derives from their exclusivity. It does not require identification of valid contributions to some universal knowledge. Consequently, new developments may be driven by myopic criteria with no assurance that they are pointing into a positive direction.

Much of the work in science and engineering is conducted within projects that are limited in time and scope. The ultimate source of power over these projects rests with institutions that provide funding. At least in the USA, these are organizations concerned with national security, high-tech enterprises, financial institutions, and billionaire entrepreneurs. All are driven by the contributions to their respective purposes. At the heart of the purpose for each is the need to succeed to defeat whoever might be their competitor. Combined with the utilitarian ethic and the notion of the invisible hand that turns bad deeds into good ones, any contribution to the growth of the overall economy provides justification for deception and dishonesty [12]. Through the words of a popular football coach, "winning is not everything, it is the only thing," it has become an integral part of the American ethic [15]. Institutions that provide funding for projects in science and engineering, therefore, will look favorably on proposals for systems or artifacts that can improve their competitive position and will actively solicit such proposals. Common to all is that they help dominate their respective domain. This includes inflicting harm on a competitor and obtaining superior information.

While the artifacts that emerge from this environment, in themselves, represent a considerable danger to humanity, a much larger danger comes from the overall system that produces them. The danger of the artifacts may be easy to spot and to mitigate. The danger from the overall system may evolve unrecognized until it is too late to mitigate it. This overall system is society (or humanity) itself. Any effort at mitigation of this threat will be futile unless some of the basic artifacts that drive society (or humanity) are questioned. It will not help to outlaw "killing robots" when "killing" is an essential function of the system. And ultimately, the system itself will turn into one big killing machine that humans will lose control of. This is not the stuff of science fiction. In fact, as the following example shows humanity has already been very close to losing control, not just of a system they created, but their own future altogether and may well still be on the way to losing control.

At the height of the cold war, the doctrine of "Mutual Assured Destruction (MAD)" was created as a system to prevent an all-out confrontation between the only two then-existing nuclear superpowers [8]. Based on the mathematical (logical) model of game theory, it was argued that if each party would always be ready to respond to an initial attack by the other with immediate and total destruction, any incentive for an initial attack would be eliminated.

Clearly, this system violates basic principles of engineering design. It is essentially unstable. Any small failure (error), rather than being contained, may be amplified and could trigger mutual destruction. (For an artistic depiction of this we refer to the memorable movie "Dr. Strangelove"[6]). Also, any small advantage by one party could be used to overwhelm the other. This provides an incentive to both parties to outdo each other in an unending arms race. Finally, rather than through a global cataclysm, this arms race may lead to the economic exhaustion of one or both parties. In the first case, both parties would have lost control of their destiny. The second case, as it happened, led to the surrender of control of the future of humanity to the remaining party. To the extent that the remaining party includes only a small portion of humanity, one may well ask what this might mean in terms of humanity being in control of its own future. In summary, MAD was a simple system that is totally human-made of logical constructs that, potentially, led to a loss of humanity over its own destiny while its ostensible purpose was exactly the opposite.

Beginning the design of an alternative system as a purposeful system does not seem difficult. However, the design process would quickly encounter issues that pose considerable obstacles to finding a feasible solution. The principle issues are human nature as well as the prevailing system of utilitarian ethics and other artifacts that derive from it. The difficulty with human society is illustrated by Niebuhr who finds that, even as all members of a society may be following moral (ethical) principles, society as a whole is bound to pursue immoral ends [9]. Ortega y Gasset anticipates the current crisis and warns that the rising importance of technology will lead to the emergence of the "mass man," a technologist who lacks the basic humanistic qualities required for responsible and a thoughtful leadership [5]. It appears that it is indeed the mass man who has led us into the current crisis. And an escape from this crisis will require leaders who, in addition to their technical qualifications, are deeply grounded in humanism. Such leaders would be able to successfully unmask the trappings of the utilitarian ethics and reestablish principles of human solidarity, and diffuse the currently prevailing combat-like attitudes. With that, it would be possible to conceptually walk back recently developed artifacts until a suitable state can be found from which it is possible to begin to chart a path back to the present and beyond that satisfies an acceptable purpose. Since this cannot be done literally, it would be necessary to develop artifacts that neutralize past developments.

5 Some Specific Issues

Since the late 1980's, the World- Wide-Web has brought about an explosive growth of instant electronic communication and information processing. In the initial phase, this growth was largely led by enthusiasts and hobbyists. Many important projects were started as informal collaborations with little or no funding. In this fluid atmosphere, there was little interest in developing strict rules that would define the boundaries of operations, or adapting existing rules from related fields. This lack of rules provided the flexibility for innovation to take hold and for small start-ups to grow to world-class corporations in a matter of years. With this world-wide reach into virtually all facets of modern life, however, the absence of rules is becoming a great concern. Creating such rules would close opportunities for the loss of control by humans. The rules listed in the following appear to be totally justifiable, and have long been proposed by others. We believe it is important to view them as essential elements to assure that humans remain in control of their own destiny.

Except as required by legitimate interests of a sovereign state, privacy is recognized as a basic human right [13]. Nobody should be allowed to enter the private spaces of any person, except with the explicit permission of that person. The private spaces include the home, cars, as well as businesses and business transactions. Many parties do have ready access to private information by the nature of the way the information is collected, stored and transmitted. The right to privacy requires that all who handle private information abstain from incidental or surreptitious collection of it. Although hacking is sometimes viewed as a test of programming techniques, it is a surreptitious entry into the private space of an individual for the purpose of causing harm. Only if privacy is assured to and by everybody will it be possible to eliminate the possibility of manipulation of individuals or groups on social media or through direct email campaigns or the use of private information for personal gain.

Another way in which humans can be left without control of their own destiny is through deception. Deception is a category that does not appear to be divisible making it necessary to exclude all forms of deception in any communication, including advertising, and the dissemination of falsehoods "Fake News". Anybody who issues (posts) news of any kind ought to be held accountable for its veracity. Enforcement of that requires that the authorship of any news item be traceable and the anonymous distribution of any kind of information be prohibited. Nobody should be permitted to issue news or statements of any kind under a false name or to participate in any forum under a false name, or create bots or filter bubbles. It should not be permitted to treat truth as a probabilistic concept that can be approached in a hit-and-miss fashion. Consequently, no one should be permitted to distribute software that potentially includes bugs.

Finally, there is the question of respect for human life. As long as killing of a human is considered legitimate, it is impossible to assure that human-created systems will not turn their killing skills against humans.

6 Conclusion

We have drawn a wide arc from systems of "mechanical" intelligence to ones that incorporate the real-life features of humans and found that artifacts that define human actions are the key to leaving humans in control. It is the human who designs a system of "artificial" intelligence and the threat from such systems can be seen as a failure of the human. A much greater threat emanates from systems that appropriate life from the humans that are included in them. It is those systems that have come closest to wresting control from the human. Reversing developments that have led to their current state appears to be of utmost urgency. These are first and foremost norms defining and guiding the behavior of individuals and of society, as well as the structures that organize human interaction and commerce.

References

[1] "Artificial Intelligence", *Wikipedia*, 22-Aug-2017.

[2] R. Brown, "US General Warns of Out-of-Control Killer Robots," *CNN*, Online, Available: http://www.cnn.com/2017/07/18/ politics/paul-selva-gary-peters-autonomous-weapons-killer-robots/index.html, accessed: 09-Aug-2017.

[3] M. Dowd, "Elon Musk's Billion-Dollar Crusade to Stop the A. I. Apocalypse," *Vanity Fair*, New York, 26-Mar-2017.

[4] Jose Ortega y Gasset, *History as a System and Other Essays toward a Philosophy of History*, W.W. Norton & Company, New York, NY, 1962.

[5] Jose Ortega y Gasset, *The Revolt of the Masses*, W.W. Norton & Company, New York, NY 1993.

[6] S. Kubrick, *Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb*, Columbia Pictures Corporation, Culver City, CA, 1964, http://www.imdb.com/titl/ tt0057012/, Accessed: 12 December 2017.

[7] Elon Musk, "Elon Musk Warns Governors: Artificial Intelligence Poses 'Existential Risk'," *NPR.org*, 2017, Online, Available: http://www.npr.org/sections/thetwo-way/2017/07/17/537686649/elon-musk-warns-governors-artificial-intelligence-poses-existential-risk, accessed: 22-Aug-2017.

[8] "Mutual Assured Destruction", *Wikipedia*, 28-Jul-2017.

[9] R. Niebuhr and C. West, *Moral Man and Immoral Society: A Study in Ethics and Politics*, 2nd Edition, Westminster John Knox Press, Louisville, KY, 2013.

[10] "Open Letter to the United Nations Convention on Certain Conventional Weapons.pdf," 2017, Online, Available: https://www.dropbox.com/s/g4ijcaqq6ivq19d/2017%20Open%20Letter%20to%20the%20United%20Nations%20Convention%20on%20Certain%20

Conventional%20Weapons.pdf?dl=0, Accessed: 21-Aug-2017.
[11] Paul Ricoeur, *History and Truth*, Evanston: Northwestern University Press, 1965.
[12] William Sweet, "Jeremy Bentham, Internet Encyclopedia of Philosophy," 2017. [Online], Available: http://www.iep.utm.edu/bentham/, Accessed: 23-Jun-2017.
[13] "Universal Declaration of Human Rights," 06-Oct-2015, Online, Available: http://www.un.org/en/universal-declaration-human-rights/, Accessed: 04-Jul-2017.
[14] N. Wiener, *God and Golem, Inc., A Comment on Certain Points where Cybernetics Impinges on Religion*, The MIT Press, 1964.
[15] "Winning is Not Everything; it is the Only Thing," *Wikipedia, the Free Encyclopedia*, 2017.

Maximilian M. Etschmaier is a native of Austria. He holds a Doctorate in Engineering from the Technical University in Graz, Austria, and an MS in Operations Research from Case Western Reserve University, where he was a Fulbright Scholar.

He is an Adjunct Professor at San Diego State University and also principal of GME International Corporation, where he has been advising clients on policy and strategy development, leading system development and process improvement ventures, and supporting international technology transfer. Previous positions include Chairman of the Management Board of Joanneum Research in Graz, Austria, Professor of Engineering at the Universities of Pittsburgh and Massachusetts, Senior Scientist at the United Technologies Research Center, Head of Operations Research of Deutsche Lufthansa AG, and Visiting Professor at the University of Graz and University of Innsbruck.

Dr. Etschmaier's professional work is focused on the design, analysis, and operation of complex systems in a wide variety of domains. The present paper reports on one of several case studies he has developed with Dr. Gordon Lee for this purpose.

Integrating Humans and Machines into Purposeful Systems that Keep the Human in Control

Maximilian M. Etschmaier*
San Diego State University, San Diego, CA 92182 USA

Gordon Lee†
San Diego State University, San Diego, CA 92182 USA

Abstract

Human-created systems are becoming ever more powerful and increasingly include the potential to harm large numbers of people, threaten their livelihood, and ultimately the survival of the human race. The vision of machines as mere tools of humans is increasingly being eclipsed by machines of such complexity that the human mental capacity no longer suffices to control a system. If humans are to stay in control of the systems they create, they need to make use of the cognitive capabilities of the very systems they are endeavoring to control. This requires an effective integration of the human and the machine. We use the paradigm of a purposeful system to re-examine the role of human insight and creativity in a system to resolve interdependence of purpose and system boundary. Since the system is a process, the human role spans the entire system life-cycle.

A number of examples are used to show how humans are properly integrated into the system and how their control over the system can be assured. The systems considered range from ones of limited scope, and including mostly physical machines, all the way to systems of potentially global impacts in which the "machines" are mostly non-physical mechanisms, including algorithms and legal and ethical frameworks. The examples confirm that the threat to human control over machines is very real and that urgent action is required to reverse the threat – if it is not already too late. We show that, through hierarchies of purposeful systems, it may be possible to design system architectures that can assure that humans stay in control. At times, this may require that, under certain circumstances, some humans, for a limited time, submit to the authority of the "machine."

Key Words: Human-machine systems, artificial intelligence, robots, social media, humans losing control, machines controlling minds, purposeful systems.

* College of Sciences. Email: metschmaier@mail.sdsu.edu.
† Dept. of Elec & Comp Engr. Email: glee@mail.sdsu.edu.

1 Introduction

It is considered an essential feature of the human race that they create and use tools to expand the range of their capabilities. While mechanical tools have expanded physical capabilities, tools for writing and painting increased human memory and made it possible to communicate across space and time. The evolution of complex social structures and societies could not have happened without these tools; neither could there be bodies of scientific knowledge or philosophical insight that depend on the confluence of the work of many individuals.

Today, with increasing power of science, including the physical sciences, information technology, control theory, computer science, artificial intelligence, the range of capabilities of tools is greatly increased. It is possible to create mechanisms ("machines") that can perform many tasks that once were reserved for humans. Machines can sense and understand much of what goes on in their environment as well as within them, and process the acquired information into decisions without human intervention. This makes it possible to communicate directly with other machines and participate in joint decision-making processes. As a result, machines are combined into ever larger complexes that operate largely without human intervention. Humans may be included in such complexes to perform functions that, with the existing state of the art in science or technology, it would be either impossible or too expensive to design machines for. In a general sense, their role may be identified as "operators."

In addition to serving as "operators," humans serve as designers, and as beneficiaries or targets of machines. The role of the human as designer is being challenged to some extent by the emergence of artificial intelligence ("algorithms") and learning systems. And the beneficiary or target may be subsets of society, or society as a whole, or some construct of nature (or the physical world) or society. This leaves the human in a very ambiguous situation.

In this paper, we examine what the role of the human might be and what needs to be done to assure that the complexes of machines of ever-increasing scope will continue to serve the

interest of humans, rather than enslave them. We believe that by viewing the emerging "complexes of machines" as parts of systems with a purpose that reflects the human interest, it will be possible to assure indefinitely the dominance of the human over "machines" that are ultimately the products of human creation.

We will follow the previously defined notion of a "purposeful system" to examine what constitutes a "system," what it is that makes a complex of machines part of a purposeful system. Clearly, a system that serves a human-defined (or human-focused) purpose needs to reflect human values and reflect the limitations of human capabilities. Behind this is an understanding of the place the human occupies in the universe. We will start with an examination of the role of humans in rather simple mechanical systems and identify how, even in simple systems, an understanding of human nature relative to the universe is of critical importance. It will become apparent that the observations derived from simple systems apply directly to highly complex machinery and, ceteris paribus, also to systems that are dominated by social relationships. Examples of the former are automated manufacturing facilities, self-driving automobiles, and computer-controlled airplanes; examples for the latter are systems for planning and operational management of enterprises, algorithms controlling social media, and arrangements to achieve global environmental sustainability.

On a practical level, we will examine what exactly is the nature of the relationship between the human and the non-human elements of a system. What is the optimal form of this relationship? How can we design systems in which the potential of the human and the machine are both used most effectively, the system does what the humans intends it to do, and meets ethical, legal and other applicable norms?

Parts of this article are based on material provided in [19].

2 The Relationship between Humans and "Machines"

This paper deals with systems that combine humans with human-created artifacts of any kind from simple mechanical devices, to complex machinery, to computer systems and algorithms, to rules governing human behavior. Ultimately, our aim is to show that to some extent all can be examined through the same logical framework. To simplify the discussion, we will refer to any type or artifact as "Machine."

To examine the role of humans in systems, it is useful to first look at the most elementary configuration of a system, a mechanical machine that is operated by a single operator and was designed and built by one single enterprise, the "designer."

A seemingly perfect integration of a human and a machine was observed many years ago at a big steel fabricating plant in Austria [10]. A pneumatic stamping machine was used to punch holes into sheets of heavy steel. The sheets were about 2'x2' and intended to serve as screens on large drums to wash raw coal. The holes were arranged in an array of 20x20. One screen at a time was mounted in a jig that could be indexed to the positions of the holes. The operator used a ratchet to move the sheet in the y-axis by one hole at a time. After a row of holes was punched, the operator unlocked to jig and moved it to punch the next row. When all holes were punched, the operator removed the sheet from the jig and replaced it with a new sheet. Each stroke of the machine was activated by the operator via a foot pedal.

This looked like a perfect arrangement for sharing work between a human and a machine. It was ergonomically reasonably well designed. And clearly, the human operator was firmly in control of the process. Nevertheless, there was something troubling about it. The plant was producing a large number of these screens, keeping the single operator working all day, for several months. There was nothing in the work of the human that would have required human capabilities. The human could have very easily been replaced by a simple mechanical device. The work was simply dehumanizing. But it is not clear that the system for producing these screens was badly designed. The operator was a poor, uneducated woman with few skills that could be used in other places of the plant. The small income she received for this work might have been the only way she could have supported herself. Whether the roles were properly divided between the human and the machine, therefore, can only be judged in the context of the socio-economic environment. Looking at the situation in isolation may lead to designs that may be well-intentioned, but totally misguided.

The next example shows more directly that the role of the human in even a simple system may be more complex than commonly assumed. We shall demonstrate this on a steam generator that was one of the cornerstones of the industrial revolution.

Steam generators (boilers) are machines that convert chemical energy into a mechanical one. In the early industrial age, they were widely used in factories to generate pressurized steam as a source of power for machinery, and to propel vehicles. Today, they are used in thermo-electric and nuclear powerplants. Steam generators are pressure vessels heated by an external heat source, in the original versions, a coal fire. The energy generated is dependent on the rate of firing which is controlled by the human operator through the volume of coal in the fire pit, and the position of the flue damper.

High demand for energy is met by a high firing rate, resulting in steam at high pressure and temperature. This in turn increases the stress on the containment walls of the steam generator. Unchecked, this will lead to an explosion of the steam generator with catastrophic results. The human will have lost control of the steam generator. To assure that the human stays in control, the steam generator is equipped with a safety valve that will stop the increase of pressure. It will open to release steam when the pressure reaches a level that approaches the strength of the walls. Control of the steam pressure is thus shared by the human

operator and the physical system; when the human might try to generate more steam than the generator can handle, the machine will take over and block him.

But who is "the machine"? Since a physical system is not capable of conscious action, it must be the human who has instructed the machine to stop the increase of pressure. This would be the designer. In other words, control of the steam generator occurs through two independent mechanisms. Both are driven by humans; one directly and in real time, and the other indirectly through the machine and at the time of design. The two are inherently in conflict with each other.

Clearly, the pressure limit set through the safety valve can be exceeded considerably before the generator will actually explode. Also, there are times when the demand for steam cannot be satisfied without exceeding the pressure limit. Sometimes the excess demand will represent a critical need that, if not satisfied, will lead to considerable harm. This raises the question of priority between the two mechanisms. Different schools of thought resolve the question differently: one assigns absolute priority to the operator who, they assume, can judge the relative importance of the two objectives, safety and demand; the other, assuming that the operator is not able to assess the complexity of stresses that result from exceeding the pressure limit, will assign absolute priority to the safety concern articulated by the designer.

The conflict between the human and the machine thus resolves as a conflict between two separate human elements, the designer and the heater (operator). It can be resolved if both, together with the physical steam generator, are viewed as being included in one system. Since the environment of the system, and the understanding of it, will evolve over the life cycle of the system, the design process cannot stop when the system is first put into service. It must continue, educated by the service experience and other learning, throughout the entire life cycle. This requires that a designer needs to remain as an essential part of the system.

The choice of the set point at which the safety valve will release steam is determined by criteria that derive from the purpose of the system. Only some of these are readily quantifiable in economic terms. Others need to be derived from nonmaterial values such as ethics and legal norms. Most prominent among them is the consideration of human life, in particular the question whether human life may ever be traded for economic benefit. Answers to this question require an analysis of the framework of ethics that governs humans.

In general, a prerequisite for the steam generator, as any system, to function as designed requires some form of sustainment (maintenance) program. Such a program will perform recurrent husbandry tasks and monitor the condition and operation of the steam generator and identify situations where a changed condition will create the risk of a failure. In the simplest case, it can be carried out by the operator. However, defining the program requires elaborate analysis, which requires access to the knowledge base that resides with the designer. The program, then, defines a complex process that requires the existence of some form of organization. A defect of the process or of the organization may cause a failure event that is outside of what should be expected; and may be catastrophic. In a recent example, a steam generator that was part of a HAVAC system exploded and thrown several hundred feet, killing four people [36]

A defect in the sustainment process or organization, thus, can be considered as a simple example of the human losing control over the system, as can be a failure in the definition of the operating procedure.

In complex technological systems, examples of humans losing control abound. Many accidents in civil aviation fit into this category. As one example, an airliner on a transoceanic flight, at cruise altitude and on autopilot, temporarily lost all three airspeed signals, eventually sending the aircraft into a stall. The pilot struggled to gain an understanding of the situation. But as he was close to gaining control of the aircraft, the airspeed signal returned briefly, leading the flight control computer to try to regain control as well, unfortunately with control inputs that were opposite to those of the pilot. The tug-of-war continued until the aircraft became totally uncontrollable and plunged into the ocean [2].

Concern about losing control over human-created systems is not limited to systems in a high technology environment. Most likely, it has existed since the beginning of humanity. When humans learned how to start a fire, they would have recognized quickly that, left unattended, fire could spread and cause considerable harm, possibly even hurting or killing the very person that started it [21, 23]. In other situations, runaway environmental damage from industrial activities remained unnoticed until the damage was irreversible, or at least catastrophic. Examples are the deforestation of mountain ranges surrounding the Mediterranean to supply ship-building. The resulting erosion of the soil was irreversible [8]. Or the unmitigated release of pollutants from steel-making until it became undeniable that the resulting pollution killed large numbers of people [26]. In all these situations, the gradual increase in industrial activities misled humans into believing that the resulting damage could be sustained forever. Clearly, humans were not aware that their activities were part of a system that also included the environment.

Even when the danger is recognizable, the desire for increased power or knowledge, or just for adventure may lead humans to ignore the risk of losing control. In a medieval story retold by Goethe, the scientist Dr. Faustus pledges his soul to the devil (Mephistopheles) in return for unlimited power and knowledge during his lifetime [43]. Tacitly, he hopes that he will be able to outsmart Mephistopheles, and his day of reckoning will never come.

Icarus is following his father Daedalus on a flight to escape from captivity on an island. Despite the strict warning by

Daedalus that the design of their flying machine requires them to stay within a fixed band of altitude, Icarus cannot resist the temptation of his curiosity to rise ever higher - until he loses control. He gets so close to the sun that his machine melts and he falls to his death [34].

According to Jewish tradition, at least since the middle ages, at various times and places, rabbis and alchemists succeed to create, out of clay, human-like figures with superhuman powers. Some, like the famous golem of Prague, were created to lead Jewish populations out of dangerous situations. Others were used to carry out commands as household slaves. These golems were brought to life through some references to god, either inscribed to them or attached. Singer [11] points out that, if a golem did possess human-like qualities, he needed to have free will. This implies that it was an essential characteristic that he could disobey, and even turn against his human creator. Given that the raw power of the golem may surpass that of the human, success in creating the perfect golem thus presents an existential threat to human existence. Wiener [44] uses the golem as a point of departure for a discussion of the relationship of the creator to his creation and warns that automation will deliver what humans design into it rather than what they meant to design. A goal-seeking mechanism will not necessarily seek our goals unless we design it for that purpose," … And "The penalties for errors of foresight … will be enormously increased as automation comes into full use."

This tradition of the golem continues through the literary works on Frankenstein [37] and the Robot [4]. Singer [38] points out how today's developments in science are bringing new relevance to the stories of golem. Modern science fiction goes beyond the notion of a human-like creature and, instead, envisions the sum of all human inventions to coalesce into some form of super-being that will control the world including the humans in it. Kurtzweil [29] and Venge [42] describe this as "the singularity". And there is some speculation that this singularity is already upon us – or at least that it is inevitable.

There are at least two ways in which humans can lose control of their destiny:

- The interaction of human-created systems with the natural system of the planet evolves into an irreversible process that diminishes the action space for human sustainment until it is insufficient to support all of humanity. At that point, humans are trapped between the human-created systems and the natural system of the planet with ever diminishing resources to support their existence.
- The algorithms driving social networks create self-contained societies, social islands that are incapable of communicating. At that point, the relationship between the islands is entirely determined by the algorithm. The human will have lost the power to control the processes that are going on between the islands. But since the algorithm will control the relationship between all humans, its dominance will also extend to relationships within an island. The ultimate fate of society will depend on the rules included in the algorithm. It may include extinction of the human race.

The human-created elements of systems encompass the whole spectrum of technological artifacts from purely mechanical devices to complex, computer-controlled machinery, to self-driving vehicles, to communication devices and social networks. But they equally include artifacts that are not related to technology and have existed since the beginning of humanity. They include language which provides the basis for human thought, learning, and all form of inter-human relations. They also include artifacts expressed in language, like laws and regulations, customs, rules of etiquette, ethics, and esthetics; also, theories of the behavior of nature; and rules and models for the organization and control of an economy.

Our contention is that all these mechanisms that are created by humans combine with humans to form systems that may leave humans in control; or they may seize control from humans and trap them in situations from which they cannot resume control. In the latter case, the result may be social structures and societies that will deprive humans of their freedom and may not be palatable. Ultimately, this may spell the end of humanity – at least of humanity the way we know it.

Many of the artifacts that may lead to humans losing control are developed without humans being aware of the full potential inherent in the artifacts. Maintaining control will then require at least a rational approach to creating the artifacts that consciously preserves ultimate control in the hands of humans. Clearly, innumerable artifacts already exist and are driving humanity in directions that may not be desirable or sustainable. These require rational examination before new artifacts can be added through a rational design process.

We suggest that a rational design process will follow the paradigm of purposeful systems to assemble new artifacts from smaller elements, all of which are themselves designed following the paradigm of purposeful systems. As shown in the next section, in theory, the paradigm of purposeful system defines a rational process through which any system can be designed to assure that the human will stay in control. However, there are essential limits to the rationality that humans can deploy. These stem from the limits on human knowledge, as well as the ability to properly use the knowledge they possess.

3 The Paradigm of a Purposeful System

The concept of a purposeful system was defined in [16] as a holistic approach to analyzing and designing any kind of human-created system. It makes it possible to consider all factors that might be of relevance to a system and recognizes the importance of human insight in its creation and operation. To this end, the system purpose and boundaries are chosen interdependently in such a way that the purpose, to the extent

possible, is included within the boundary. The limitation comes from the fact that only the universe could completely contain its purpose.

The purposeful system is a process, a dynamic construct that continuously adapts as the environment changes and knowledge about it evolves. The human element forms an integral part of the system. In addition to the humans that perform operational roles, the purposeful system also includes the "historian" and the "designer," both of which remain with the system throughout the entire life-cycle. Their role is to continuously monitor and analyze the performance of the system and to modify the design as emerging information about the system and its environment dictates. This is done by continuously recreating the design process.

Both, the "designer" and the "historian" are human or include human elements. As that, they are subject to the limitations of knowledge of reality spelled out in the next section. Although their design and analysis are guided by subjective judgment, the resulting knowledge becomes objective fact [20, 35]. In engineering and business systems, the role of historian may be assumed by what is commonly referred to as a learning function, provided that the scope of such a function is not limited to automated analysis but includes a human analyst who is free to engage in the required introspection.

The biggest innovation that the paradigm of purposeful systems provides is through the way the design process starts. Unlike engineering design, which is commonly defined to start with a statement of a problem that the design is expected to solve, the process of designing a purposeful system starts with some seed, which is just an idea, or a spark of intuition of the designer. This idea is then translated into some initial sense of purpose, followed by an exploration of functions that might be required of a system that can meet the purpose. This then starts a circular process that iterates between reexamination, and presumably expansion of the statement of purpose, and of a system boundary that can envelope the collection of functions until the purpose can be enclosed by the system boundary.

This is an essentially creative process that is driven by the "designer," who can be an individual, or a team, or an organization. Etschmaier et. al., [42] and [15] have demonstrated that the format of a "functional system diagram" may be an effective means to use during this process and to document the result. However, since this process is based on the insight, intuition, and imagination of the human, there can be no rigid rules that would restrain it. This will assure that the system that is being designed is free of prejudice and, while guided by experience, will not be constrained by preconceived notions. It will provide the intellectual freedom necessary to break through perceived barriers and abandon old stereotypes, such as what is commonly perpetuated by rigid rules and procedures.

The result of this process is the first step of system design. It is a functional model of the system. In the second step, a hierarchy of components is identified to which the functions can be attached. Again, this is largely a process that is driven by human creativity and insight. Etschmaier [17] proposes an object-(agent-) based representation through which a purposeful system can be defined in terms of objects (agents) and functions in such a way that the system purpose can be pursued in the most effective and efficient manner, and unacceptable system states, such as catastrophic failures, are avoided. The model includes all the information necessary to monitor and control the system operation as well as to manage the process of evolving throughout the life cycle of the system. It can be used throughout this phase of the design process as well as through the construction, operation and sustainment of the system.

A conceptual representation of the life cycle of a purposeful system is shown in Figure 1. The system evolves along three axes: real time, the state of system creation, and the system creation (evolution) time. Within this space the system moves as an ever-increasing plane spanned by the two axes system creation time and state of system creation in the direction of time. This plane is where the continuous analysis and redesign happen by circling through the steps of the original creation, or more precisely, continuing the circular design process. The upper vertex represents the evolving system state. This is the locus where the system delivers on its purpose. This is where system control, sustainment and monitoring occur. It is where the system design process and system operation intersect.

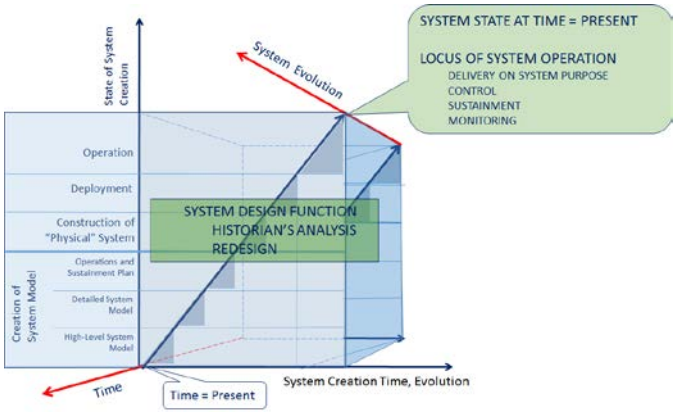

Figure 1: The life cycle of a purposeful system

The purpose in a purposeful system is an integral part of the system. As the purpose is defined by humans, humans essentially are parts of any purposeful system. Conversely, due to the consciousness of the human, all systems created by humans can be regarded as purposeful systems. Clearly, this applies to engineering systems. It also applies to social systems. Even systems that are not meant for any purpose other than play and diversion can be regarded as purposeful.

In a way, the process of designing a purposeful system can be seen as mimicking the process of creation of life. The beginning

of the process is where some initial spark happens in the designer's mind - the big bang. And with this spark, the system can evolve. But it is not through the creation of new "matter," but through spanning a net over existing objects ("matter") and including them within the boundary of the system by providing them with meaning (purpose) as part of the system. And this process will shape each object, both physically as well as ideally (logically). Including an object in the system does not necessarily require removing it from other systems. And it is this membership in multiple systems that creates the tension of any system with "the environment". It is a circular process through which the system evolves.

We suggest that the process of designing a purposeful system is how any good engineer has always designed new systems, especially systems that brought about radical change to processes and to industries [11, 12]. It is also the process through which radical change can be developed in social systems. Following Thomas Kuhn [28], that is also the way in which scientific knowledge evolves and "scientific revolutions" happen.

Clearly, this design process is inspired by the way operations research work was defined and conducted in the early days, at Case Institute of Technology, as the scientific approach to tackle and improve real-life situations [1]. Unfortunately, Operations Research soon became too fascinated by mathematical methods through which models could be developed that approximate real situations. These models, then, were used to draw conclusions about real situations, and to develop prescriptions for designing and operating systems. Eventually, the process became reversed. A repertoire of models was developed as the standard tool set. And models were applied even to situations where they did not provide a perfect fit. Reality became viewed through the lens of available models. In many situations, that, in turn, changed reality. The well-known book "Factory Physics" discusses this process and the resulting problems at length [22]. Most recently, Operations Research has undergone yet another transformation, replacing human insight by algorithms ("data analytics") that work on extremely large data-sets to develop insight into real-life situations and how to deal with them [24].

The paradigm of a purposeful system encapsulates the process of designing the system. [17] shows that the design process is essentially the same as the process of analyzing a purposeful system. While the design process may be viewed as creating a system where there was none, the process of system analysis attempts to identify a purposeful system that fits into an existing situation, in other words, explains some reality in terms of the paradigm of a purposeful system – and identify changes that would or could bring an identified purpose within identified system boundaries. Thus, what may have started as an effort of analysis, merges into a design process.

Applied to social systems, the paradigm of purposeful systems might well be considered as an addition to the analysis tool set of Sociology in general, and in particular the field of Science and Technologies Studies [5, 30]. The object-based structure of models that the paradigm lends itself to can be helpful in the development of quantitative simulations. However, it can also be configured to handle analyses and simulations of conceptual and logical constructs.

4 Limits of Knowledge of Reality Defines the Structure of Purposeful Systems

The question of our relationship to reality has occupied humanity at least since the time of classical Greece. Plato famously concluded that humans lack a direct access to reality. All we can perceive of reality is a projection onto some screen, for example, the wall of a cave. Over the years, discussions among philosophers led to questioning the existence of anything that we perceive as real, and eventually our very existence. The pronouncement "cogito, ergo sum" by Descartes elegantly put an end to this uncertainty. But the question of how our perception relates to reality remains.

Most recently, neuroscience is providing physical evidence that vision, our visual perception, is only a phantasy that our brain is creating on the basis of surprisingly little information combined with past experience [25]. Proving this point, a rich set of experiments shows convincingly that the brain can predictably be fooled. For example, a brief video proves repeatability that under certain circumstances an object, in this case a gorilla, can be moved across the screen without being noticed. In fact, it leaves observers convinced that there was no object [40].

Thomas Kuhn [28] shows that our relationship to scientific truths is not much different. His theory of scientific revolutions shows that the scientific community embraces a dominant theory even as evidence emerges that it does not completely explain reality. And the majority of scientists (representatives of "normal science") remain committed to the theory even as a new theory ("scientific revolution") emerges that can better explain reality.

In popular conception, engineering systems consist of physical components that interact with each other and with the environment following precisely defined rules, physical laws responding to inputs from human operators and the environment – predictable, like clockwork [9]. Deviations from the rules occur due to wear, a predictable process, or due to mistakes of the operator. Clearly, no system ever followed this description. Any component of a system always included variations in material or dimensions that could not be controlled, i.e., were essentially random. In the steam generator example discussed above, the rate of firing could only be controlled in very crude terms, the strength of the boiler walls or of the welds varied widely, hopefully within some given tolerance limit. Beyond that, we now understand that the laws of Newtonian physics that were assumed to govern engineering systems were themselves only an approximation of reality. Today, as the physics used in

the design of engineering systems has moved a long way beyond Newtonian physics, we recognize that understanding the true state of nature remains as elusive as ever.

Thus, even in engineering systems, we do not have access to what the actual system is. We are limited to seeing the system through a model that we ourselves build, a phantasy. And this model is not the same as what the designer intended to design, which in turn is different from what he actually designed. While these differences may be small, they may be quite significant because different models may project different system behavior. Thus, there is significant uncertainty for the human element in the design, construction, and operation of a system. Humans encounter the system from different perspectives, each through a different model, and each assuming they are dealing with the same system. Effective control of the system requires harmonization of the different views.

One way of achieving convergence is to invest in gathering and processing of information to bring the different human-created models closer to agreement with the underlying reality. The investment required is different for each model and yields a different contribution to improvement of overall system performance. Not every investment actually contributes to improved performance. Rather, it is necessary to view all investment together and search for a balance. Such a balance may exist for many levels of investment. It represents a "consistent set of information" for the given level of investment and the optimal interaction between the human and the physical parts of the system. This approach leads to a static optimization of the system.

Another way to achieve harmonization of different models is through dynamic monitoring of system behavior from the point of view of different models. Convergence may be achievable by looking for differences in predicted system behavior by different models as they are exposed during system operation. Figure 2 shows a schematic of how this can be accomplished. The scheme is based on the insight of the designer and the historian that is an essential part of a purposeful system. The historian monitors and analyzes the system performance for differences to what is predicted, as well as for emerging differences to the underlying reality. Based on this analysis, the designer will identify and implement improvements to the system model.

Figure 3 shows that in a realistic system there are a multitude of models that are at different distances from the true state of the system. In the figure, the closeness to the true state is increasing from right to left, as indicated by increasing density of the color. This means that the model of Figure 2 will have many dimensions and the relationships will be defined along a complex network of interdependencies. While it may be difficult or unnecessary to actually develop a graph of this network, an exploration of the relationships in one way or another will play a central role in designing a purposeful system.

The following example shows how the understanding of

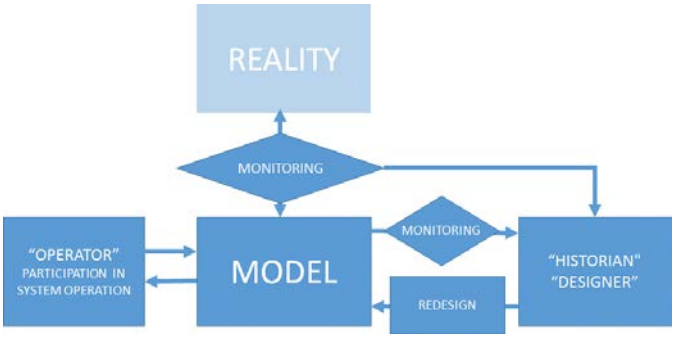

Figure 2: Human interaction with the system only through a model

Figure 3: The multi-layered relationship between human understanding and true laws and state of nature

reality can evolve naturally within an enterprise. This was the great innovation in airliner maintenance that took place in the 1960's that made modern, inexpensive, and ubiquitous air travel a possibility [33]. Unfortunately, at that time the full depth of the approach was not adequately described and the underlying concepts are no longer understood.

At the time, airliners had been maintained according to fixed schedules, pegged to operating hours, frequency of certain stress conditions, or elapsed time. At or before defined instances, predefined work packages had to be performed on the airframe or components, such as, in particular, engines. The aircraft could not continue flying until the work was completed. Statistical analyses were used to determine rates of deterioration. Simple cost-benefit analyses were used to determine the "optimal" interval between work packages. Additionally, repairs were performed as demanded by failures.

The point of departure for the development of the new approach was the observation that many components, especially engines, that were disassembled for maintenance on schedule showed little deterioration that would have justified performing the scheduled work. Worse, faulty work caused new failures. At the same time, it was recognized that many failing components emit a variety of signals of their deteriorating state. These signals can be monitored and used to determine when a failure is becoming imminent. When work is called for, it can

be tailored to what is actually demanded by the specific state of the aircraft or a component.

What is called "maintenance program" is the framework through which signals from components are monitored and analyzed and through which human intervention in system operations occurs. It provides the human with the information he needs to recognize an imminent (catastrophic) failure, initiate the collection of additional information and the performance of analyses to find the intervention that is most effective and efficient. It addresses system processes that result from wear and tear and from impacts of the environment (both predictable and unpredictable), and from inadequacies in the design process. The inadequacies can be the result of human error, or aspects of system behavior that are beyond the state of the art. The maintenance program is designed to assure that none of these will lead the system to enter an unacceptable (catastrophic) state, and to correct non-critical failures through measures that are most effective and efficient.

Monitoring the condition of the aircraft and its components is distributed over all humans who have contact with the aircraft and have occasion to notice changing conditions. This limits the effort for collecting the information, but spreads responsibility for sustainment of the aircraft across the organization. This requires that humans across the organization are always in a state of high alert and that they are embedded in a well-functioning organization that promptly responds to signals demanding maintenance work.

The process of creating the maintenance program gives rise to examining the system design and the possibilities for humans to understand it. Specifically, it may identify instances of inadequate design where the aircraft is exposed to catastrophic failures that cannot be prevented by any maintenance activity or other human intervention. For the aircraft to be fit for use, it has to be redesigned. Similarly, targeting maintenance work to actual deterioration requires analysis that can identify design features that makes the aircraft inherently unsafe to operate and require correction of the design.

In summary, different individuals throughout the organization may see the aircraft and its condition through different models. Examples, in addition to maintenance crews, are cockpit crews, cabin crews, and ground crews. The model for each one of them will need to evolve over time in order to keep the aircraft "alive."

5 Human Capacity to Function as Part of a Purposeful System Derives from Insight Provided by Access to the Evolutionary State of the Noosphere

The value of the human element in a system derives from the cognitive ability, ability to invent, create, and think outside established categories. These are not qualities that are unique to each human. They all draw from the same source, the global consciousness of humanity. This is what Teilhard de Chardin [7] named the "Noosphere," an entity that forms the top rung on the ladder of evolution. It is the result of humanity reflecting on itself. In it any human thought becomes an entity of its own and merges and competes with other entities, forming the leading edge of evolution of the universe. From this evolutionary process emerge powerful movements that drive the thoughts of humans. The movements interact, coalesce, and compete with each other. It is only at the end of time for which Teilhard de Chardin envisions convergence to one singular point.

This provides one model for why some individuals approach their role in a common system from the same perspective, and others from different ones. The implications of this for the design of a human-machine system can be considerable. It impacts the design as well as options available for human control of systems, what humans can be expected to know, how they can be expected to think and evaluate situations, how they can be expected to act. The models they work with can be influenced by purposeful or subconscious actions such as education, indoctrination, literature, art, sports events, and propaganda.

A similar model is due to Durkheim [5], who is considered to be one of the founders of modern sociology. Durkheim views a society as an "ensemble of ideas, beliefs, and sentiments of all sorts that is realized through individuals …" And "… society and social phenomena can only be explained in sociological terms as the fusion of individual consciences that, once created, follow their own laws."

The field of organizational behavior is trying to bring about a certain amount of harmony among all members of a system. Training and educational programs try to harmonize the technical competence and establish "standards of work." In the end, it is essential to recognize that different individuals are driven by different models which all originate from a powerful source. A human-machine system (or in fact any system) needs to be designed in such a way that such differences, which may lead to failures of the human element, will not cause an inacceptable state of the system.

6 Failure of the Human Element

Failures of aircraft can lead to consequences that are significantly worse than those of many other types of equipment. As a consequence, a rigorous safety culture has evolved that is supported by strict laws and regulations. Serious accidents are examined by teams of experts to find the cause of, and ways to prevent similar accidents. Frequently, the cause is identified as "pilot error," a failure of a human component of the system. Usually there is a chain of events that puts the aircraft into a compromised state from which the aircraft might have recovered if the pilot had intervened decisively with a set of actions that were taught in flight training, spelled out by the operating manual, or could have been gleaned from the context described in the manual. In other cases, there is an error or an

omission which leads to a compromised state from which the pilot cannot recover the aircraft. Only rarely is there a situation where the pilot willfully puts the aircraft into a compromised state from which it cannot recover.

In all these situations, it is assumed that the pilot is the ultimate authority for controlling the aircraft. Often it is assumed that the pilot could have acted differently and thereby avoided the accident. Even as pilots often do act heroically to recover from a compromised situation or to mitigate the consequences if full recovery is not possible, this cannot be assumed as the norm and expected of every pilot. Especially in aircraft that include a high degree of automation, it may be next to impossible for the pilot to quickly understand all relevant variables of the state of the aircraft, diagnose what went wrong, and identify suitable actions to prevent an accident. It is the model of the pilot as the supreme authority of the aircraft that is failing.

It can be recognized that the pilot is only one component of the system aircraft, and that the system "aircraft" as a whole has failed. The fault, then, lies with the design of the aircraft. This is the responsibility of the designer who handed over responsibility for the aircraft to the operator at the point of delivery. The design and the operation of the aircraft are parts of different systems – systems that actually have competing purposes.

The concept of a purposeful system puts both the design and the operation of the aircraft into one system. Such a system will overlay parts of the entity that designed and supplied the aircraft and the airline that operates it, while otherwise maintaining the independence of the two intersecting entities.

Viewing the situation as a purposeful system makes it possible to limit the authority of the pilot while maintaining the overall dominance of the human element. It is just that the ultimate human authority over the system has been moved to the designer, who is supported by the historian. This makes it possible to justify arrangements that are already commonplace, like restricting the control input to keep the system within a safe operating envelope. The equivalent in the example of the steam boiler is to limit the pressure of steam in the boiler even when the heater would want a higher pressure to satisfy an operational need. For realistic examples, we refer to several recent papers [13, 17, 20].

A purposeful system includes functions (control structures) that protect the integrity of the system, i.e., prevents the loss of functions that would enter the system into a prohibited state. This applies equally to functions that are carried out by physical components as well as by humans. This means that the design has to be inherently safe and the system remain safe throughout the entire life cycle. Catastrophic events can only be prevented through proper system design. Vulnerabilities in the design are uncovered by monitoring for deviations between system behavior expected from the model with actual system behavior. The deviations are analyzed for indication of impending catastrophic failures. Safe emergency shut-down procedures need to be included in the design.

Only events that are caused by circumstances outside the state of the art are really unavoidable. To the extent possible, the design can allow for their occurrence and provide for mitigation of the consequences. Other causes are human errors in the design processes. They can be dealt with in the design in the same way.

7 Keeping Humans in Control of a System

Humans can lose control of a system in many situations that vary widely in the scope of their consequences and in the time scale in which the consequences manifest. In the following examples, we shall review cases that range from ones that impact a small number of humans in a period of minutes to ones that have impact on a society and happen over years.

In the Air France accident introduced above [2], the pilot and the flight control computer were fighting each other. While there is some speculation about who was right, in the end, the human lost (or did not regain) control of the aircraft because the aircraft plunged into the ocean. How could this be avoided? The accident investigation identified a number of events that, together, made the loss of the aircraft unavoidable. But at the heart of it is the design of the system that manages (controls) the flight. Viewed as a purposeful system, this includes, as components, the flight computer, the pilot, the cockpit resource management (CRM) procedure that defines rules for the interaction of the two pilots between each other and with all information resources available to them in critical situations, and all potential sources of information. In a purposeful system, all components would be examined for the functionality they could provide to assure continued safety of flight. It would not be possible that the flight control computer would simply turn itself off when it recognizes a faulty input (in this case from all three airspeed sensors, and attitude information that puts the aircraft outside the envelope of feasible states). And it could not simply turn itself on again when it resumes to receive plausible input. Instead, it would recognize the potential for critical failures in both transitions and present to the pilot whatever information it recognized as trustworthy, in a format that can readily be used by the pilot to avoid a critical failure.

Etschmaier and Lee [17] analyze the case of the Malaysia Airlines Flight 370 that stopped communicating with air traffic control and is assumed to have continued flying an unknown course over the ocean until it ran out of fuel. They show that, irrespective of what actually happened on the aircraft, a purposeful system design could prevent the human from losing control. The purposeful system, for this purpose, in addition to the aircraft and cockpit crew, needs to include the air traffic control system (including meteorological information available through it) as well as information about the status of all reachable airports. The human who is left in control, though,

may not be the cockpit crew, since they must be assumed to have failed or may be incapacitated by terrorist action. It may not even be any person who can directly impact the flight path, since any one of them may be subject to coercion by terrorists. Instead, it needs to be a person (or persons) who is removed both in space and in time. That person is the designer, who can include in flight management a procedure that will autonomously direct the aircraft to a safe (emergency) landing if it recognizes that the aircraft is on a trajectory that rules out the possibility of an orderly completion of the flight. It is interesting to note that in this situation real-time control may be taken away from humans in order to preserve overall control of the aircraft by humans.

The case of self-driving vehicles requires a much larger perspective. Self-driving cars do not constitute entirely new systems. Rather, they represent an augmentation of the capability of existing systems. Before examining how humans can maintain or lose control, it is necessary to examine the situation with conventional cars.

Looking at individual cars, we can clearly see that humans lose control of them all the time, either on their own, or together with others. If ethical norms demand that human-designed systems do not harm humans, 35,000 traffic fatalities per year in the US alone clearly show that humans have long lost control of the system (automobile) [31]. It is because we assume the fact that a human is at the controls of the car means the human is actually in control that we do not consider this situation as objectionable as it is.

Replacing the human by the equipment to make the car "self-driving" does not make the situation any less objectionable, even if it brings about a significant reduction of traffic fatalities. It would be a false choice to ask a self-driving car to decide who to kill and whose life to spare in a traffic accident. Clearly, in a system that cannot altogether avoid harming humans, the human has lost control. The fact that this will require that cars travel at a lower speed than is currently the case is not important. Actually, under current law, accidents should not be possible either. The law clearly states [3] that the driver should not exceed a speed that would enable him/her to safely stop in front of any obstacle on the road. The system of posted speed limits does not supersede this rule. Unfortunately, it confuses drivers who, misled by enforcement practice, tend to assume that speed limits are the only rule.

The sensors included in self-driving vehicles could easily support a driving regime that follows the law. But such a self-driving vehicle would not fit into current flow of traffic. By slowing down for hazards, it would probably be considered an obstacle. However, if posted speed limits were eliminated (as they could since the car can by itself determine the safe speed) higher speed at the open road might more than compensate for this. This means self-driving vehicles could indeed put humans back in control of traffic if some relatively simple changes are made to the current practice of managing traffic. These changes would not require a separation between conventional and self-driving cars; but would require changes in the enforcement of traffic laws.

However, there is another dimension in which the automobile has caused the human to lose control, and may already have lost control. That stems from the impact the automobile has had on society, the economy, and in particular land use and the environment. It may well be that the patterns established by the introduction of the automobile are not sustainable in terms of burdens on the environment, wasteful use of land, or impact on social hygiene. The environmental resources consumed by developed countries may well exceed what would be their share if the developing countries could attain the same standard. The recently concluded global agreement on climate change mitigation [4] shows how difficult it is for humans to regain control of just one part of the problem automobiles are responsible for. The other parts of the problem have yet to be addressed at a global level.

Civilian drones are remotely controlled, partially autonomous aircraft that are rapidly gaining in importance. It appears they are being built because it is possible to build them. And only now is there a debate over how they can and may be used, and how to regulate use. Issues involving their design, their use, and their operation should have been recognizable before the first drone was built – and could have been dealt with then. In many ways, issues are similar to those of automobiles. However, management of three-dimensional space is vastly more complex. And while drones do not have drivers or passengers to worry about, there is the possibility of a mishap that could send a drone falling to the ground and cause harm to humans and to property. Additionally, through their ability to enter spaces that are otherwise inaccessible, they provide new venues to harm humans physically, or in terms of deprived privacy. Keeping humans in control requires mechanisms to prevent harm that results from accidents as well as from terrorist intent. Similar to the automobile, the use of drones may usher in developments in the economy, in society, and in the environment. Especially for the long term, these mechanisms may well limit the use of drones. Now would be the time to study and develop a purposeful system that includes the universe of drones and includes functions to eliminate the possibility of harm to humans and to avoid irreversible developments for the long-term.

While military drones might appear to be similar to civilian drones, their purpose as remote killing machines is turning them into something altogether different. Current generations of drones come equipped with a killing device, usually a weapon, and a set of sensors to find and home in on prospective targets. One or more operators, possibly a continent or more away, will identify a target, direct the drone into a suitable position, direct the killing device at the target, and survey the resulting damage. Although firing deadly weapons from a long distance away has long been part of the craft of warfare, there are two features that

set drones apart. The extreme distance from which drones can be controlled eliminates any risk of harm to the operator; and the "surgical precision" of the killing process. Chamayou [6] shows that both features have already fundamentally changed the nature and rules (and laws) of warfare. Any country that can gain a dominant technological advantage over all other countries can project its power to all parts of the world and subject the entire planet to its rule. Following Niebuhr [32], power in that country will invariably tend to converge into an ever smaller group which will exert its power in an immoral manner. It may be speculated that such a group will ultimately self-destruct. At least at that point, humanity will have lost control.

Current developments are aimed at truly autonomous drones that, once launched will be controlled by an algorithm to select, find, and destroy a target. As Wiener [48] said in 1964, these drones will do what they were told, but that might not be what humans wanted to tell them. Like Hal in the Kubrik movie 2001: A Space Odyssey [27], the drone might turn against its creator. Extrapolating from that, the fear is not unfounded that autonomous drones might start a war against humanity and wipe out humanity, the ultimate form of loss of control by humanity over its creation.

Election systems are one of the most important tools in the operation of a democracy, providing the citizens the mechanism for selecting representatives in the political arena, and for expressing their opinions on certain issues of importance to them. They comprise systems to record and accumulate votes, to aggregate votes at the local, regional and national levels, and to keep a registry of eligible voters. A failure of any one of these systems may inflict serious damage to the state, and literally make a difference between war and peace. Failures may happen by accident, through the failure of any one of the functions of the system or through intentional manipulation of any of the system components.

The most vulnerable part of the election system is the system to record and accumulate votes. This used to be a ballot box and the vote count by a team of election officials. Today it is mostly a "voting machine" of one of several designs, generally an electro-mechanical or electronic device. It is vulnerable because the law requiring ballots cast to be secret means that there is no way to verify proper operation.

Research by Etschmaier [18] has shown that this problem can be overcome through design, construction and operation of voting machines as a purposeful system. He showed that it is possible to develop a generic design in the form of a government regulation, following which specific designs can be realized in any technology. This is in contrast to current regulation which covers specific technologies, is based on encryption, and verification of compliance through testing. Encryption is an arms race with "hackers," and testing can only verify that a particular specimen of a voting machine has performed properly during the test. The experience with emission regulation of automobiles shows that testing is anything but reliable. The fact that the operation of a voting machine cannot be observed during the actual election process makes tests even more uncertain than is the case with automobile emissions.

The experience of the election of 2000, where the outcome of the presidential contest was literally dependent on the confidence in the performance of voting machines in one county in Florida, as well as of several elections since, arguably demonstrated how just uncertainty about the results can seriously damage confidence in the democracy. Unless concrete evidence of manipulation of vote counts emerges, the true result will never be known. And the damage from the lingering uncertainty will only grow, a process the human (at least the voters) has lost control of to a defective election system.

Etschmaier has shown that loss of control can be avoided if voting machines are designed, constructed, and operated following regulation based on the paradigm of purposeful systems and are imbedded in an election system that is designed following the same paradigm of purposeful systems.

On the surface, social media provide platforms through which individuals can communicate with, or present their views to one or preferably more partners on a regular basis. Since all information conveyed through communication includes a certain amount of opinion, the effect on the partner will include some amount of persuasion.

Unfortunately, the information users reveal is valuable to the host of the social medium, especially if is stored over a longer period of time. It can be analyzed using complex algorithms ("big data analytics") and collated and correlated with data from a variety of other sources to produce detailed profiles of users, often including information the user is not, or not yet aware of. Marketers can use the profiles to precisely target advertising messages. Crafted using a rich repertoire of psychological and communication tools, these messages are designed to shape the user's preferences, to create desires for goods that he/she does not otherwise need or want; and to subconsciously implant opinions in the user's mind. The same tools can be used to promote certain political arguments, and to persuade individuals to participate in certain events or to join certain movements.

The host of the platform will also try to promote the use of the platform, increasing the relevance of the platform by connecting users with similar views. The long-term effect is the evolution of "filter bubbles," groups of users who primarily communicate with each other, amplifying each other's opinions as well as their opposition to other opinions. In the end, the population will wind up divided into almost completely disjoined groups with ever stronger attachment to the platform.

Since this process increases user loyalty and keeps attracting ever more users, it increases revenue for the platform. It is the secret of success of the platform and is responsible for the astronomical valuation of some platforms. Inherent in this is that virtually all revenue of the platform derives from taking away control of the users over their own mind. The platform is surreptitiously taking control over the users, sweetening the

experience by providing some utility to them.

That taking control over users' minds is a real threat can be seen on the most recent US presidential elections. This may be much more of a threat to the system of democracy than manipulation of the election system. In a way, social media may be viewed as tools to manipulate election outcomes. The disparity between polling results and election results may be an indication that social media have gained unprecedented power. A particularly stark scenario of what might have already happened is presented in [39]

In summary, social media, by their very nature, are designed to take control over the user's mind. Leaving the user in control would mean reducing the platform to a means for users to exchange information that the host neither reads, nor processes in any other way, nor stores.

Social media can also be manipulated by users to amplify their own opinions and messages. They may insert "bots," a form of robot simulating a human to amplify the momentum behind their message. This is not against the interest of the host, even if the opinions violate social norms, or are used to form or promote criminal associations. It may be possible for government to craft laws and regulations that prohibit exchange of illegal content or the use of bots. Alternatively, deleting offending messages by the host would quickly run up against laws guaranteeing freedom of speech. Anyway, either approach could not change the process that is essential to the business model of the platform, to take control of users' minds. This model is not only reaping enormous profit for the host of the platform. It is also contributing its entire revenue to the growth of the economy. Curtailing that would violate the central ethical principle that is currently driving the economy, the notion of the invisible hand that turns even ill-gotten economic gain into a positive contribution to the economy. Any government that would attempt to curtail the success of a social-media platform would be seen as hurting the interests of the state, violating the commitment it made to the nation.

Thus, it follows that it is the economic system, a human-made construct of principles and rules, that is directly responsible for humans losing control of their own mind. Ironically, the underlying ethical system ("philosophy"), utilitarianism, ardently promotes the freedom of the individual. Any attempt to return control over their own mind back to the humans would threaten the very principles the state, through its economic system, is based on. Speculating how such a feat could be accomplished transcends the bounds of this article. We end by pointing out that the paradigm of purposeful systems possesses the power to identify the contradiction within the current principles governing society, and to develop approaches to remedy the situation.

8 Conclusion

The concept of a purposeful system views a system as a dynamic construct that evolves through continuous adaption throughout its entire life cycle. Different from the prevailing notion of a system, there is no point in the life of the system when an "operator" takes over full control. Instead, the "designer" and the "historian" remain as integral parts of the system to guide its evolution throughout the life, particularly to assure that the system is kept from entering prohibited (catastrophic) states. The holistic concept of a purposeful system thus makes it possible to recognize the complete role the human element plays in a system.

Critical to the definition of the role the human element in the system is the appreciation of the fact that the design of any system is based on a limited understanding of the true state of the system and its environment. Humans can see reality only through models. It is essentially impossible to create models that are completely congruent with reality. However, the degree of similarity between the model and reality can be increased through targeted acquisition of information. Over the life time of the system, additional information can be obtained through monitoring the difference between the system state predicted by the model and what can be observed of the state of the system. Analysis of that difference is what drives the continuous redesign of the system and what provides the human operator with tools to steer the system clear of prohibited states (catastrophic failures).

The process of system design and redesign is based on humans identifying possibilities of events that could drive the system to an unacceptable or prohibited state, and designing mechanisms to guard against such events. The object-based model structure of a purposeful system makes this possible. This is significantly different from alternative approaches that are based essentially on statistical analysis, such as risk analysis or big data analysis.

We have attempted to show that the notion of purposeful system can be used to model any type of system that includes human made mechanisms along the entire spectrum, from mechanical apparatuses to laws and regulations, and to fundamental worldviews the state is based on. If used judiciously, it might help remedy some of the contradictions (and confusion) that are currently emerging in the operation of the commonwealth. We believe that we have also demonstrated that the power inherent in what a good engineer has always done is not limited to the creation of physical systems but can successfully address social issues and paradigms of the mind that determine the course of the world.

References

[1] R. L. Ackoff, *The Art of Problem Solving*, John Wiley, New York, NY, 1978.

[2] "Air France Flight 447," *Wikipedia, the Free Encyclopedia*. 06-Jun-2015.

[3] "California Driver Handbook - Laws/Rules of the Road,"

2017, Online, Available: https://www.dmv.ca.gov/portal/dmv/detail/pubs/hdbk/speed_limits+, Accessed: 08-Jun-2017.
[4] Karel Čapek, *R. U. R.* Pocket Books, 1970.
[5] P. Carls, "Emile Durkheim," *Internet Encyclopedia of Philosophy*, 2017.
[6] G. Chamayou, *A Theory of the Drone*, The New Press, New York, NY, 2015.
[7] T. de Chardin, *The Phenomenon of Man*, Harper and Row, New York, NY, 1959.
[8] "Deforestation during the Roman Period," *Wikipedia*. 25-May-2017.
[9] E. Dolnick, *The Clockwork Universe: Isaac Newton, the Royal Society, and the Birth of the Modern World*, Reprint Edition, Harper Perennial, 2012.
[10] Maximilian M. Etschmaier, "Observation of a Worker at Stamping Press at Alpine Zeltweg," Personal Notes, 1957.
[11] Maximilian M. Etschmaier, "Relationship between Automation and Economy under Consideration of the Human Factor," *Plenary Address*, Computaport 84, International Civil Airports Association, Tel Aviv, 1984.
[12] Maximilian M. Etschmaier, "Institute for Engineering Leadership," College of Engineering, University of Lowell, 1984.
[13] Maximilian M. Etschmaier and Gordon K. Lee, "Designing, Analyzing, and Operating Secure Systems," *Proceedings*, 29th *International Conference on Computers and Their Applications*, Las Vegas, Nevada, pp. 107-113, 2016.
[14] Maximilian M. Etschmaier, S. Rubin, and Gordon Lee, "A System of Systems Approach to the Design of a Landing Gear System: A Case Study," 27th International Conference on Computer Applications in Industry and Engineering, New Orleans, LA, 2014.
[15] Maximilian M. Etschmaier, S. H. Rubin, and Gordon K. Lee, "On the Use of SOMPA Core Modeling for Systems Design: a Case Study," *Proceedings*, Kona, Hawaii, 2014.
[16] Maximilian M. Etschmaier, "Purposeful Systems: A Conceptual Framework for System Design, Analysis, and Operation," *Int. J. Comput. Their Appl.*, 22(2)87-100, June 2015.
[17] Maximilian M. Etschmaier and Gordon Lee, "Defining the Paradigm of a Highly Automated System that Protects Against Human Failures and Terrorist Acts, and Application to Aircraft Systems," *ISCA*, 23(1):4-11, Mar. 2016.
[18] Maximilian M. Etschmaier and Gordon K. Lee, "Designing Secure Computer Systems as Purposeful Systems," *IJCA*, 23(2):105-115, Jun. 2016.
[19] Maximilian M. Etschmaier and Gordon K. Lee, "Purposeful Systems - Truly Integrating Systems of Humans and Machines," *Proceedings 32nd International Conference on Computers and Their Applications*, Honolulu, Hawaii, pp. 155-160, 2017.
[20] Jose Ortega y Gasset, *History as a System and Other Essays toward a Philosophy of History*, W.W. Norton & Company, New York, NY, 1962.
[21] J. A. J. Gowlett, "The Discovery of Fire by Humans: A Long and Convoluted Process," *Phil Trans R Soc B*, 371(1696):20150164, Jun. 2016.
[22] W. J. Hopp and M. L. Spearman, *Factory Physics*, Third Edition, Waveland Press, Inc, Long Grove, Ill, 2011.
[23] "Human Nature, Technology & the Environment," 2006, Online, Available: http://fubini.swarthmore.edu/~ENVS2/S2006/seberha1/prehistoric.htm, Accessed: 08-Jun-2017.
[24] INFORMS, "Operations Research & Analytics," *INFORMS*, 2017, Online, Available: https://www.informs.org/Explore/Operations-Research-Analytics, Accessed: 08-Jun-2017.
[25] E. Kandel, *The Age of Insight*, Random House, New York, NY, 2012.
[26] Edwin Kiester, Jr., "A Darkness in Donora," *Smithsonian*, 1999, Online, Available: http://www.smithsonianmag.com/history/a-darkness-in-donora-174128118/, Accessed: 08-Jun-2017.
[27] S. Kubrick, *2001: A Space Odyssey*, Metro-Goldwyn-Mayer**,** 1968.
[28] T. Kuhn, *The Structure of Scientific Revolutions*, Second Edition, The University of Chicago Press, Chicago, 1970.
[29] R. Kurzweil, *The Singularity Is Near: When Humans Transcend Biology*, Penguin Books, New York, 2006.
[30] B. Latour, "On Technical Mediation - Philosophy, Sociology, Genealogy," *Common Knowledge*, 3(2):29-64, Fall 1994.
[31] "List of Motor Vehicle Deaths in U.S. by year," *Wikipedia*. 26-May-2017.
[32] R. Niebuhr and C. West, *Moral Man and Immoral Society: A Study in Ethics and Politics*, 2nd Edition. Westminster John Knox Press, Louisville, KY, 2013.
[33] F. S. Nowlan and H. F. Heap, *Reliability-Centered Maintenance*, Dolby Access Press, San Francisco, CA, 1978.
[34] Ovid, *The Metamorphoses of Ovid*, 1 Edition, Harcourt Brace, 1995.
[35] Paul Ricoeur, *History and Truth*, Northwestern University Press, Evanston, 1965.
[36] Emily Shapiro, "3 Dead after Boiler Explosion in St. Louis," *ABC News*, St. Louis, 03-Apr-2017.
[37] M. W. Shelley, *Frankenstein*, Prestwick House Inc, 2005.
[38] Isaac Bashevis Singer, "THE GOLEM IS A MYTH FOR OUR TIME," *The New York Times*, 12-Aug-1984.
[39] "The Great British Brexit Robbery: How Our Democracy was Hijacked," *The Guardian*, 07-May-2017.
[40] "The Invisible Gorilla: And Other Ways Our Intuitions Deceive Us," 2010, Online, Available: http://www.theinvisiblegorilla.com/ index.html, Accessed: 08-Jun-2017.

[41] "The Paris Agreement - Main Page," 2016, Online, Available: http://unfccc.int/paris_agreement/items/9485.php, Accessed: 08-Jun-2017.
[42] Werner Venge, "The Coming Technological Singularity: How to Survive in the Post-Human Era," VISION-21 Symposium Sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute, March 30-31, 1993, 1993.
[43] Johann Wolfgang von Goethe, "Faust", *Goodreads*, 1988, Online, Available: http://www.goodreads.com/work/ best_book/16721-faust-eine-trag-die, Accessed: 08-Jun-2017.
[44] N. Wiener, *God and Golem, Inc., A Comment on Certain Points where Cybernetics Impinges on Religion*, The MIT Press, 1964.

Maximilian M. Etschmaier's professional work is focused on the analysis, design, and operation of complex systems in a wide variety of domains. He is currently an Adjunct Professor in the College of Sciences at San Diego State University. Previous positions include, Guest Researcher at the National Institute of Science and Technology (NIST), Senior Scientist at the United Technologies Research Center, Chairman of the Management Board of Joanneum Research in Graz, Austria, Vice President of Systems and Control at Bricmont Associates (now Andritz Bricmont), Professor of Engineering at the Universities of Pittsburgh and Massachusetts, Head of Operations Research of Deutsche Lufthansa AG, and Visiting Professor at the University of Graz and University of Innsbruck. He has advised business and public-sector clients on policy and strategy development, lead process improvement ventures, and supported international technology transfer.

Dr. Etschmaier is a native of Austria. He holds a PhD in Engineering from the Technical University in Graz, Austria, and an MS in Operations Research from Case Western Reserve University, where he was a Fulbright Scholar. He has participated in national and international scientific and professional organizations, serving in leadership positions, organizing and hosting meetings and sessions, and presenting and publishing numerous papers.

Gordon K. Lee was born and raised in Hawaii. He received his B.S. degree in Electrical Engineering from the University of Hawaii in 1972, his M.S.E.E. degree from the University of Connecticut in 1974 and his Ph.D. degree from the University of Connecticut in 1978. From 1978 through 1989, Dr. Lee was at Colorado State University in the Department of Electrical Engineering where he rose to the level of Full Professor. He was also the Director of the Institute for Robotic Studies.

In 1989, Dr. Lee became a faculty member in the Department of Mechanical and Aerospace Engineering at North Carolina State University and also served as Director of Graduate Programs in the Department of Mechanical and Aerospace Engineering and later as Assistant Dean for Research Programs in the College of Engineering. Dr. Lee joined San Diego State University in December 2000 where he served as the Associate Dean and Director of the Joint Doctoral Program for the College of Engineering. He was also a full Professor in the Department of Electrical and Computer Engineering and is currently Professor Emeritus in that department.

His research interests are in the areas of robotics and intelligent control systems, particularly evolutionary control algorithms, fuzzy systems and neural networks, as well as in the applications of these methods to mobile robotic colonies. His research projects have been funded by government agencies as well as industry. He has published over 275 technical documents; Dr. Lee is a senior member of IEEE, a member of AIAA and a senior member of ISCA. He is also currently an Associate Editor for the International Journal on Intelligent Automation and Soft Computing.

Beneficial AI? Fight for It!

Eugénio Oliveira*
University of Porto, Porto, 4200-465, PORTUGAL

Abstract

An "Artificial Intelligence-First" world is being preached all over the media by many responsible players in economic and scientific communities. This paper presents our belief in AI potentialities while warning against the current hyping of its near future. Although quite excited by the interesting revelations of several recent books like "The Master Algorithm", we try to argue in favor of a more cautious interpretation of the AI-based systems and algorithms potential outreach. This does not prevent us from discussing the possibility of approaching some kind of consciousness through Artificial Intelligence, arguing against well-known Searle's former statements on the subject. We also include in the paper some personal perspectives on simple remedies to preventing recognized possible dangers. We advocate a set of practices and principles that may prevent the development of AI-based programs and systems prone to be misused for the harm of humans and raise some ethical issues we believe should be discussed now. Accountable "data curators", appropriate software engineering specification methods, the inclusion, when needed, of the "human in the loop", software agents with emotion-like type of states contributing to the reasoning process are important factors leading to more secure AI-based systems. To inseminate ART in Artificial Intelligence, ART standing for Accountability, Responsibility and Transparency, becomes also mandatory for trustworthy AI-based systems.

Key Words: Artificial intelligence, beneficial AI, machine learning.

1 Introduction

Most relevant outcomes of civilization derive from intelligence. Heliocentric, evolutionary and relativistic theories are examples of how brilliant scientific minds intelligently reshape the way humans dramatically changed their perspective about the world. How can we improve and enlarge those benefits, through artificial intelligence (AI) based systems, without being fully replaced both on the job market and, most important, as final decision-makers? AI has evolved, during the last five decades, starting with a very classical approach grounded on mathematics and psychology and followed by more romantic times in which almost everything was said to be possible. "Within a generation [...] the problem of creating 'artificial intelligence' will substantially be solved," Minsky is quoted as saying in the book "AI: The Tumultuous History of the Search for Artificial Intelligence" [4]. By the end of the seventies of the last century, the AI community was more directed towards a pragmatic stage where it was considered mandatory to make impact on society by solving realistic problems. Instead of teaching computers to solve logic puzzles or play chess, the AI pioneer Edward Feigenbaum, in the seventies and eighties of the last century, at Stanford where he developed the concept of expert systems, urged AI researchers to get out into the real world and solve real-world problems [12]. This approach, although leading to real useful applications in the Knowledge Based Systems field, was followed in the nineties by a call back to the fundamentals in which learning, adaptation, cooperation and autonomy became corner stones of intelligent systems. It was not very long ago that a rupture in the step by step intelligent systems development happened and, together with euphoria, new warnings reached the scientific community about the future potential dangers of "AI winter" has already showed us what bad consequences has resulted from AI being over-hyped in the past.

This paper presents our belief in AI while warning against the current hyping of its future. It includes our perspective on the potential risks of "general AI" developments, including learning machines, as well as some personal perspectives on simple remedies for preventing a few of the recognized potential dangers. It finally raises some ethical issues that, in our opinion, should already have been taken into consideration.

2 Can Artificial General Intelligence be Dangerous?

AI-based systems have been performant and useful in many different relevant, although narrow, domains. For example, these systems have been used, to make specific medical diagnoses, to allow companies to build up consumer profiles, for satellites to be intelligently controlled, for search engines to do page ranking, and for computers to intelligently filter spam. A recommender system such as those used by Amazon and Netflix are welcome. And we feel proud of amazing accomplishments of AI programs like the ones that employed by Deep Blue and AlphaGo. This corroborates what R.

* Faculty of Engineering and LIACC-Artificial Intelligence and Computer Science Lab. Email: eco@fe.up.pt.

Brooks said in 2006, when he opposed the idea that AI had failed and warned that AI would be around us every day (reproduced in [22]).

Artificial general intelligence can be seen as an intermediate stage between what we have now, a kind of artificial specialized intelligence that is very performant in restricted domains, and a conceivable future superintelligence that might, endow artificial systems with the capability to exceed human performance in many, if not all, the relevant domains, possibly including leadership. Because most AI-based systems, in some way, reason and interact, we are often tempted to compare them to humans. We sometimes forget the limitations that still make a great difference.

"*While humans are fast at parallel processing (pattern recognition) and slow at sequential processing (logical reasoning), computers have mastered the former in narrow fields and are superfast in the latter. Just as submarines don't swim, machines solve problems and accomplish tasks in their own way*." [13].

Moreover, according to some scientists and opinion-makers, we could expect that superintelligence or general intelligence, would give artificial systems the property of consciousness, making the boundary between humans and machines, in many decisive aspects, fuzzy.

Some authors [17] are now asking the following question: Is the human brain the only system that can host a mind? If digital minds come into existence, and the referred author states that it is difficult to argue that they will not, we have to face all the legal and ethical implications of such a possibility.

It is argued that current hardware sophistication and development rate, regarding miniaturization and integration, makes us believe that in a few years it will be possible to replicate the number of synapsis happening at the brain level. I believe that reasoning patterns of high level of abstraction as well as structured knowledge are not always directly emerging from those simple operations. It is however worthwhile to prepare ourselves for this future possibility. Legislation and ethical principles may guide a harmonious development of either some kind of digital minds or even hybrid minds.

It is not yet the case that we foresee the possibility of humans becoming obsolete in too many situations, but it is the right time to clearly state that real beneficial AI must be developed in such a way that humans and machines cooperate to solve complex problems together and, in doing so, possibly learn from each other. More than having intelligent entities, robots, systems, computers, machines, programs, replacing humans everywhere, we need to develop processes, methods and regulations leading to a harmonious coexistence of both for humankind beneficial. This ultimate goal justifies that we must pay attention to present signs that point to possible dangers in some future research directions of AI, leveraged by a plethora of books and scholarly opinions over-hyping the current and future role of AI.

Although it may seem so, I am in no way against the scientific development of the artificial intelligence field. On the contrary, I stand for a firm position about the crucial importance of the field. I even believe that, despite how AI, in the past, in most scholarly curricula as well as fora, has been seen as a sub-area of computer science, it is now the moment to look, at least partially, at computer science and informatics, as scientific disciplines that should study and develop computational methods to automate and process information (which is what the term "informatics" means) to build up really intelligent, and human compatible systems. This would fulfil the main objectives that are proposed by the scientific field of Artificial Intelligence. To understand the complexity of all the different facets of intelligence and to make computer systems behave accordingly should be the main goal of computer science.

Security and privacy, data integrity, distributed and parallel computation, software engineering development methods and many other computer science topics should have in mind the needs of intelligence-based systems.

Although this can be prone to controversy, computer science and informatics should thus be seen as contributing to the broader field of artificial intelligence. An artificial intelligence confined by ethical principles for research and development.

3 "Strong AI" and "The Master Algorithm" Claims

"The Master Algorithm" [9] is a remarkable book that makes us think and exercise our critical opinion without denying both the beauty and the dangers of its main message.

"*Our goal is to figure out the simplest program we can write such that it will continue to write itself by reading data, without limit, until it knows everything there is to know*."

To be able "*to know everything*" could be in itself potentially dangerous, but things still change for the worse when the same author also claims that "*Machine learning is remaking science, technology, business, politics, and war ...*", [9].

Although this last claim may be accepted as partially true, it also reveals a well-known tendency to oversell a specific research topic, trying to ignore that, often, machine learning (ML) algorithms work together with a multitude of other different algorithms in order to get things done. For example, when saying: Google's self-driving car taught itself how to stay on the road and no engineer wrote an algorithm instructing it, step-by-step, how to get from A to B, it seems that it is all about Machine Learning. But, this is not the case. There is also a need for, at least, competences on advanced computer vision and systems control, trajectory planning, sensing and perception algorithms. Moreover, you also need computer systems' distributed architectures and modules (or even software agents), and coherent interaction and coordination. Artificial intelligence should be neither glorified nor blamed in isolation for the important outcomes to appear soon.

It is true that ML algorithms look like artefacts that produce new artefacts. In some way, a "Master Algorithm" would be a powerful and absolute general-purpose learner, a kind of "Holy

Grail" which, in reality, I believe will be very difficult to find. If it exists, the master algorithm, seen as a combination of current ML algorithms working over big data, "*can derive all knowledge in the world - past, present, and future - from data*". Inventing it would be one of the greatest advances in the history of science. It would be, as the author names it, the "*ultimate learning machine"*, [9].

Science is mostly based on observations, gathering relevant data, and inferring appropriate models in which available data fit. Thus, it would seem perfectly reasonable to argue that ML over big data will enlarge the scope of science and will give us unlimited knowledge. We, nevertheless, should not forget the burden of dealing with permanently changing streaming input data and its needed pre-processing. However, it definitively seems to me that, up to now, those algorithms work over data that, although collected in large amounts, have a relatively simple or already known structure. You do not need much extra knowledge to build up a theory that explains those extracted patterns.

This is not the case whenever big data has to be first recognized and then extracted from many image-based sources (video, pictures, other images) in which recognizing what is data also becomes a crucial issue. A priori knowledge to guide the system focus of attention on different and dynamic situations becomes of utmost importance for collecting the relevant data.

Even so, amazing outcomes already confront us with ethical issues. As an example, Lyrebird developed by PhD students at University of Montreal, by applying a deep learning algorithm can, precisely enough and in one minute, generate an imitation of specific human voices. "*Lyrebird claims it can generate 1,000 sentences in less than half a second using GPU clusters.*" [5]. Impersonating someone at a distance, your boss, your relative, becomes now, more easily to achieve.

Without our explicit consent, there are also large data brokers that collect, analyze and sell to others all the harvested details about consumers' online activities for marketing purposes. Despite some obvious potential advantages, in my opinion, being guided in our so-called preferences for the sake of envisaged future activities like, for example, literature reading, political voting or wine drinking, may be indeed a bad idea since it looks like you are being held by your hand like a child. It may even be the case that, who knows, whenever you decide to act differently from what was expected, when you are upset with your past choices and decide to do it radically differently, it may happen that you will become suspect to someone or some organization, seen as a disruptive person, halfway to become a potential terrorist... On the other hand, there are complex problems for which this kind of approach will not be enough to infer useful knowledge. Since you have huge amounts of data available about climate all over the world for many decades, why is climate for the next month still so hard to preview?

Maybe it is because there is a need for more sophisticated reasoning over the big data that goes beyond those patterns that can be directly extracted. It may be that other kinds of knowledge, different from what is directly extracted from the specific data, becomes relevant for the correct interpretation of the phenomena. Really trustworthy climate models need to be derived not just directly from weather closely related data, but also from other may be already available, priory knowledge that may give new perspectives on that data interpretation. Here, both humans and intelligent machines have a lot to progress.

Are current AI algorithms ready to derive all possible and needed knowledge from any kind of data sets? Of course not. You may supply hundreds of thousands of medical cases about, let us say, different cancer types, but if you miss a few tenths of cases regarding very specific situations, they will always remain invisible to the inferred algorithms. Sundar Pichai, chief executive of Google and an AI enthusiast assures that "*Google is going to be AI first*". Very recently he even stated that "*In an AI-first world, we are rethinking all our products*" (see The New York Times, May 18, 2017). Although he is confident that AI will make available a general tool designed for general purposes in general contexts, he also adds, and I fully agree, that "*for the moment, at least, the greatest danger is that the information we're feeding them* [AI-enhanced assistants] *is biased in the first place*" [15]. Data Curators become then necessary to guarantee that the recorded past is not adulterated and remains trustworthy.

ML was an AI important research topic that steadily grew, since the seventies, mainly driven by research on symbolic learning. Other emergent different approaches with the same goal, automatic learning, were always more or less undervalued as not belonging to the same ML-AI tribe.

Connectionist and evolutionary-based algorithms have often been seen by the former ML researchers as proposing research directions waiting for their respective dead-end. ML researchers even saw themselves like the elite of AI, working upon the only topic that, in fact, they believed deserved to be considered as doing real AI.

It is amazing that now, not only ML is claiming to be "the" AI but also it is willing to encompass all the other approaches to automatic learning. Precisely this is predicted in the above referred book, the "Master Algorithm" as the result of combining the five machine learning paradigms and schools that currently exist: the symbolists, connectionists, evolutionaries, Bayesians, and "analogizers". Despite the fact that existing learning algorithms, although illustrating brilliant ideas, are, in fact, extreme simplifications of brain machinery and evolution laws, the author believes that the current state of the art is as close as anybody has come to the definitive paradigm shift towards the "Master Algorithm". In our opinion, nothing proves or justifies beyond any doubt the belief that, exactly now, we have come to the situation where we recognize that have all the bricks needed to build a solid staircase leading to the universal learning capability.

Chaining and mixing those existent different machine learning principles, may not be enough to solve the overall learning problem. Even if we accept the inherent power of data, it might take more than collected observations to directly induce natural selection "*as Darwin did*" [9]. Is it just a matter of observing data? I do not believe it was only that. Notice

that many people, many minds, and all along many, many years were not (and are not) getting everybody to the same conclusions even in the presence of the same available data. Maybe this is because we need still more than simple data to work out sophisticated conclusions. Which points out to what is missing here could be the most important element: Some kind of ability that some minds (and brains also) have developed, and others did not, to extract from as well as apply to, the same data in some identified contexts, more sophisticated knowledge. And perhaps, there are many different capabilities that need to be developed in the future that even the most gifted minds and brains cannot yet imagine.

We should also be cautious about the scope of AI and ML. In the same book it is stated that "*The Master Algorithm would provide a unifying view of all of science and potentially lead to a new theory of everything.*" [9].

I recall that, for example, a theory of everything is sought because quantum physics only deals with the very small. Einstein's general relativity theory deals with the very big and we are looking for a unique theory that works everywhere. However, physicists do not think that the theory of everything will come out of a kind of combination of the previous two theories mentioned before. They are still looking for something radically new. The same will happen, in my humble opinion, with the so-called "Master Algorithm" and it is an over simplification to believe that it (like a kind of "master key") will come precisely out of the ML algorithms that we already know nowadays. Therefore, statements like the one referred to above are, in my opinion, clearly inadequate regarding the current state of the art. I am not as radical as those who state that "*big data is not the new oil; it's the new snake oil*". But, nevertheless, I would be more cautious in targeting the possible goals of current ML algorithms working over big data as the "*ultimate learning machine*". Learning is becoming the hard kernel of AI, enabling more sophisticated and general-purpose AI-based systems capabilities. Artificial General Intelligence can be seen as fostering the property of consciousness. This property can also be translated as self-awareness or even capability of feeling (sentience).

Moreover, the well-known and typical Turing test seems a bit outdated since one of the most important skills for an intelligent agent (and system) is the capability to work in cooperation together with other either artificial or natural agents according to normative and ethical rules in order to solve complex problems. This points to the need of a new kind of test reflecting more of a sociological flavor that, consequently, could be called a "Durkheim Test" (referring to the French sociologist Emile Durkheim).

John Searle, in his book "Minds, Brains, and Programs" [20], clearly states that. "*A program cannot give a computer a mind, understanding or consciousness regardless its intelligence.*" The main argument he used, the well-known Chinese room, seems more like a paradox which, like the Zeno paradox, contradicts observed events. This is the opinion of Jean E. Tardy, who in the book "Meca Sapien Blueprint" [22] argues that machine consciousness is feasible. In fact, dividing time into episodes and analyzing their respective properties. For each partial distance, it seems impossible that Achilles overturns the turtle or, in the Chinese room example, each machine program step, at a time, only manipulates a set of symbols and does not fully understand Chinese may lead to the false conclusion that the overall machine behavior will never be able to really understand Chinese. Therefore, according to Searle, such a machine will never be endowed with consciousness equivalent to the conclusion that Achilles will never overturn the turtle.

Jean Tardy observes that it would be as if looking at separate static frames would lead us to the conclusion that these very same frames, seen sequentially would be incapable of generating movement. It is also possible to say that there is some kind of self-awareness and consciousness whenever a program, hardware-embodied or not, can establish its own goals even in a narrow sense and make its own decisions. But this would be like guiding a missile to hit a wedding ceremony in Afghanistan (suspicious set of guns recognized at the entrance of a large tent) or selling shares at the stock exchange once the market situation seems favorable.

In fact, in order to display some kind of goal directed behavior and interfere with our feelings, intelligent systems need not include the consciousness property. Anyway, this kind of debate lacks rigor since "consciousness" is still an ill-defined concept, the same applying to awareness.

"*Consciousness is equal to that specific capability also called sentience* [capable of feeling] *and self-awareness*" [22]. As a definition it does not help much. Is awareness the acknowledgment of self? How to define the self? Even if we admit that it might be possible that some simple type of "consciousness" will emerge from very complex interactions of more primitive forms of intelligence included in AI-based systems, we cannot assure that such a complexity will be reached with current "in silico" hardware systems. Moreover, the possibility either to download a mind or to make it evolve from a simpler digital mind, and, here, I agree with Arlindo Oliveira in his recent book "The Digital Mind", [17] would need a non-existing reverse engineering capability of the brain or, for the latter alternative a kind of real body, plenty of sophisticated sensors which is not yet available today.

We are aware of the efforts being done, for more than a decade, by neuroscientists and other researchers to better understand neural activity at the brain level.

The Blue Brain Project aims to build comprehensive digital reconstructions (computer models) of the brain, which include the brain's different levels of organization and their interactions, and which are compatible with the available experimental data. However, according to [11] "The project's current reconstructions reproduce the detailed cellular anatomy, connectivity, and electrical behavior of a small part of the neocortex of young rats (about one third of a cubic millimeter of cortical tissue)". Indeed, to replicate "in silico" what exists "in vivo" in the human brain it is (fortunately) out of our grasp as far as we can now preview based on scientific grounds.

4 What Can We Do?

It is obvious that there are potential applications in which data collection, data mining and ML outcomes become not at all crystal clear and may lead to conclusions backing some kind of artificially justified dominance in many different aspects. Stating that "[computers] they'll even guess what we want before we express it." [9] is the first step to impose to you what someone believes you should want according to their own, those who implemented, or payed for, the program decision-making strategies. Taking the AI researchers' role, we are here mainly concerned with establishing a set of practices and principles that may prevent the development of AI-based programs and systems prone to be misused for the harm of humans. And the first major concern is privacy.

Many data mining algorithms rely in analyzing sensitive personal data including individual identification, photos, genetic and medical records. It is not always the case that the inferred results, through appropriate generalization, fully protect users' data privacy. There are large data brokers like Acxiom, ID Analytics, CoreLogic, and Datalogix which gather, process, analyze and then sell billions of pieces of information about consumers' online activities supposedly for marketing purposes. We must enforce and support all the efforts trying to ensure that individual privacy will always be guaranteed and are not just feeding someone else's commercial interests.

Are we not over-reacting? Should we really be afraid of some potential future AI-based systems development directions? Haven't we always known how to deal with similar possible threats?

There are, in fact a few ready-to-use naive answers to those who are already alarmed with possible dangers. They could be (based on Stefan Wess) [22]:

- We do not need to be afraid of these artificial systems. Likewise, with other previous technological threats, we always can "remove the plug" and that is it.

  Well, we should not be so sure about that possibility. Many sophisticated programs and huge amounts of data tend to be distributed in the cloud and are not easy at all to reach.
- Always provide a kind of "kill switch". Do not forget that we certainly would expect that intelligent machines will, in the first place want to preserve themselves and be immune to such drastic killing possibility (through cloning, for example).
- Put the machine into a "cage". That is what a virtual machine is intended to do to protect computers from incoming programs. But, remember, this so-called solution is not safe enough even to prevent many already existent smart viruses.

We have then to recognize that the problem is real and we, as researchers and developers, need to take actions to reinforce AI-based systems security well beyond simplistic solutions. Individual privacy should not be for sale, especially by others.

5 The Human in the Loop

A few years ago, I and two former PhD students of mine design a so-called autonomous software system that was able to propose solutions for unexpected operations plan disruptions in the context of Airlines Operations Control [2]. Each different software expert in dealing with and trying to solve a specific aspect of the problem from aircraft landing delays or aircraft unavailability due to malfunctions, to crew members absence, start to look for the best solution for the problem in hand. Several different autonomous agents of the overall multi-agent system worked together and collaborated in finding a good solution to the problem minimizing the effects of the unexpected event.

Our multi-agent based solution was not looking for some kind of optimality but instead, for an explicit compromise taking into account the somewhat different interests of the participant entities (company, passengers and crew). Rather than operational research type of algorithms, a negotiation-based approach, using different heuristics, was put in place to reach acceptable solutions. However, we soon came to the conclusion that the way we built the so-called autonomous system could lead to a biased solution, even though that solution would be considered as "sub-optimal." And of course, according to that solution, the airline company would always appear to be the winner. But what about the legitimate interests of the crew members or the real individual interests of the passengers? It was not a difficult problem to find out new weighted solutions that would take into account all different perspectives. They were not as much appealing to the company as the first one but, nevertheless, they could be accommodated together with marginal impact.

The real fair and final solution calculating a combined utility taking the different perspectives into account was only achieved when we included the human in the loop principle giving some authorized officer the responsibility, without escape, of explicitly weighting those several different perspectives in order to find out what could be, according to his decision, and under his responsibility, the best compromise for the ultimate policy justifying the specific inferred solution. In some contexts, passengers, or even crew members, may be more important than the immediate and direct company interests. And someone had to be responsible for that choice. The human took the responsibility for instructing the automatic system in giving one solution more prone to satisfy the company, the passengers or the crew. Human-machine collaboration became transparent and useful.

Moreover, software agents behind the scene also included a learning capability trying to improve each time the way they negotiate with others to make their proposal better accepted by the other ones. Curiously enough, the learning algorithm we were applying does not perfectly fit in the five learning algorithm tribes classification proposed in [9]. Since the system needed to incrementally learn with very few examples, we were using a reinforcement "Q-learning" algorithm. Later, the "human in the loop" component became again a corner

stone of another semi-decentralized multi-agent system we have designed. This time the objective was managing ship damages when they are under severe conditions, either weather conditions or external attacks. In these kinds of scenarios, all the monitoring capabilities and solution-finding (plan of actions as well as crew and resources allocation, tasks prioritization) came out of the automatic software agents' decision capabilities, including the very same learning algorithm using reinforcement learning. However, it was mandatory that at least some part of the command chain was replicated for interfering with the decision system at different levels all along the decision process.

One cannot forget the intrinsic responsibilities assigned to humans (commanders and officers in the first place) in charge for the sake of the acceptance of the proposed solutions at different moments in time. Some identified individual must be accountable for the most important decisions taken. This was indeed a relevant factor in the possible acceptance of the semi-automatic solution for delicate and sensitive problems like the one of managing a ship in harsh situations. To make this possible in a transparent way, developers need to take these interactions into consideration from the beginning. Therefore, the system specification stage becomes crucial in guaranteeing that we can trust the system. Concerning good practices for AI systems development, it is wise to follow clear and well-founded specification methods.

Not only Software Engineering tools are available to ensure programming according to the previously identified and intended requirements, but also there have been for some time new interesting improvements towards the intelligent (mostly agent-based) systems development cycle starting from requirements extraction, design specification, implementation and finally validation phases. Agent Oriented Software Engineering methods [2, 24] help specify intelligent software agents, recognizing their hard and soft goals, identifying actors with their respective roles and pointing to mandatory human interaction situations. It is a good principle that we should be forced to use to be on the safe side regarding future AI-based systems behavior.

Despite a good specifications practice, is it a definitive answer to AI and ML potential dangers just to include the human in the loop? It might not be. We should not forget that "Drones can fly autonomously with the help of learning algorithms; although they are still partly controlled by human pilots" [9]. And, despite being monitored by humans, we should not be sure of the drone's goodness in many different situations.

6 Is Rationality Mandatory?

The recent western economic crisis made many economists to believe that it is wrong to build strategies upon computer-based models in which agents are believed to always act rationally. Real intelligent agents, in order to be included in economic models should be aware of more sophisticated decision-making capabilities that go beyond strict economic rationality.

In a different scientific domain, back in 1997, I published a short paper about "Robots as responsible Agents" [17]. My naive approach, twenty years ago, was when the then novel cognitive software agent architecture was based on "mentalistic" concepts like "Beliefs", "Desires" and "Intentions" (BDI) could bring a positive influence in the designing of more self-aware robots controlled by those BDI software agents. I was proposing a two-layer architecture, using symbolic representation for dealing with knowledge and goals at the deliberative level and sub-symbolic neural networks for implementing specific behaviors at the more instinctive reactive level.

One of the main problems we were addressing was how to make these two control layers to communicate, to interact and to cooperate without being completely dependent from each other which, if it were be the case, could lead to unsolvable dead-locks. Regarding intelligent robots, I still doubt that with the hardware limitations at that time, and current capabilities, we could make them evolve for a much more intelligent-like kind of entity. However, at least we could combine the two different levels of decision making, one relying on some kind of instinctive reactive capability and the other displaying a more cognitive-based intelligent behavior. While the former level should be implemented in a sub-symbolic way through artificial neural networks type of algorithm, the latter was based on the already mentioned BDI architecture. The main and most difficult issue became then, how to make those layers to smoothly work together.

Planning, learning, classification, intentions-guided decision-making capabilities should in certain situations (I would say in most of them) take control of the intelligent robot. But in other specific scenarios we may expect that reactive behavior leads to the best decisions for the sake of survivability and efficiency. The problem is that, in real scenarios, a vast grey area exists where both reactive and cognitive capabilities can be invoked, may overlap and even compete for the robot's control.

In order to make the Robot control system more responsible, taking decisions humans could better understand, we have also proposed the use of a modal logic (intentional logic) to correctly define what could be considered as persistent goals for a software agent (controlling a robot) to pursue and the conditions for giving them up avoiding getting stuck trying to fanatically reach impossible goals. My real implemented mobile robot was never able to satisfactorily solve this kind of schizoid behavior for some particular situations in which reasoning and reacting were both, simultaneously, of paramount importance. It is not here the place to go into details on this problem. It was only about five years later that I realized that one important and decisive component of human-like reasoning is deeply related with emotions and could be helpful for intelligent AI-based systems.

Some, like John Searle [21], arguing through an article in the Wall Street Journal against real intelligence of IBM Watson, the program that brilliantly won the "Jeopardy" competition against humans, sarcastically said that the referred sophisticated program did not become happy after winning. I,

nevertheless, believe that it would not be very difficult to program Watson in such a way that, after winning the game, it would reach an "emotional state" similar to happiness. Not regarding the external signs of happiness, which would be too easy to implement, but in which concerns the internal reasoning capability changes along with its way of acting and memorizing for a certain period of time until that emotional state gradually declines.

Contrary to what many scientists and philosophers advocated in the past, human emotional states have been recognized as essential for human reasoning and decision making. Neuroscientists of the last decades proved that reason and emotion are intrinsically intermingled [6] and thus, computer scientists in the quest for real artificial intelligent entities should take this relationship into account. In collaboration with other researchers, I have tried to give a contribution to logically define an emotions-based BDI agent architecture making it possible to make decisions while also taking some primitive emotions into account [18]. To have an individual perception of the personal risks any situation involves, as well as knowing about its own individual capabilities it might have available for dealing with such situations, are crucial factors influencing an individual (software agent, robot, person) final decision.

Past experiences, in different scenarios and with different meanings, can be mapped to a kind of primitive emotions (fear, anxiety, ...) intensity, through accumulator-like variables. Accumulators gain some amount of energy whenever specific identified stimulus happen sometime in the past and they discharge their energy, during a certain period of time, following specific decay curves. Including these "emotion-like" states in the reasoning loop, makes it more difficult to take decisions that possibly lead to bad results in terms of causing harm or some kind of pain to the agent. This implies that artificial and, let us say, intelligent decision-making may benefit in taking into consideration these more human-like influential factors, like emotion states, in order to become more human friendly and compatible.

The proposed initial emotional-BDI agent's architecture included in our agent-based system was endowed with ЄBDI Logic [18]. ЄBDI Logic includes a set of axioms about beliefs, desires, intentions, resources, capabilities and primitive emotions like "fear" and "anxiety". With this enhanced architecture, agents were able to recognize threats representing either facts or events occurring in the environment, which could directly affect one or more fundamental desires of the agent, thus putting at stake its goals and self-preservation. Dangerous threats, i.e. those that agents believe will potentially lead to the falsification of at least one of their fundamental desires would trigger, through the emotional-state driven engine, decisive decisions that would eliminate such recognized threats.

7 Ethical Issues

I believe we do not want to see the boundaries between the individual self and artificial systems to dissolve. Are we ready to accept what the author of "The Master Algorithm" book said in a TEDx talk: "*the question what means to be human will no longer have an answer. But maybe it never did.*"? (in "Next 100 years of your life" [10]).

Also V. Dignum defines a super-intelligent system "*as a system that exceeds the capabilities of humans in every relevant endeavor, can out-manoeuvre humans any time its goals conflict with human goals*". And warns against the possibility that such superintelligence "*may enable outsmarting financial markets, out-inventing human researchers, out-manipulating human leaders, and developing weapons we cannot even understand*", in "Robots and AI are going to take over our jobs! Or work with us for a better future?" [8].

Are we going to leave AI plus IoT (the internet of things), plus ML, to create some kind of future dystopia? Or will we be able to circumscribe the potential dangers and fortunately live with the obvious advantages of this new technology?

It seems that there is now a main concern of AI players starting with researchers and ending with big high-tech companies leading to the searching for ethical laws that could prevent situations like those happening during the industrial revolution or even those years immediately after the development of nuclear energy.

It is true that the nature of problems raised by "strong AI", "artificial general intelligence" or "deep learning" possibilities are not just of ethical nature. There are also concerns about an obvious economic impact. It is true that past threats like the nineteenth century industrial revolution showed that, with time, societies progressively adjust their economic development by creating new, sometime not foreseen, opportunities.

However, it is also a fact that, if a radical transition is made without humanistic and enlightened vision, many concrete people as well as entire economic sectors and geographical areas, will never satisfactorily recover from the technological shock. But that is another level of discussion outside of the scope of this paper.

According to J. Tardy [22] Superintelligence might not be perceptible as such. It could be diffused and with a fluctuating identity. It may even be the case that the so-called "technium", the network of all the computers, devices and things (to use a concept bought by Arlindo Oliveira from Kevin Kelly in [17]) becomes a synthetically controlled highly distributed and heterogeneous service network, responsible for a non-localized set of events. This possibility emphasizes the actual worry that AI could turn competent but with goals misaligned with well-formed human goals [22]. That is why so many people are now contributing to the discussion on how to guide future AI research development in such a way that, whatever results we will get in the future they point to a beneficial AI age.

We stick in line with the 23 Asilomar principles pushing AI research towards the creation of, not undirected intelligence, but beneficial intelligence instead [1]. However, these principles respective analysis is out of the range of this paper. We are also aware of the efforts made by M. Delvaux, at the European parliament, about the possibility to give intelligent

robots a limited "e-personality", that could be comparable with what already happens with "corporate personalities", a legal status which enables to sue or to be sued in court. If we have learned something from the past about law, it is that it does not change as fast as technology does. We will have to wait a long time before relevant legal system changes will occur.

We thus prefer here to emphasize that we should enforce decisive principles to be applied to AI systems, like those brought from good corporate governance and that again V. Dignum [8] also advocates: To inseminate ART in Artificial Intelligence. Here, ART stands for Accountability, Responsibility and Transparency. We need to know, in all circumstances, who is to blame whenever an AI based system's misconduct is noticed, the typical example being the situation of a self-driving car accident harming pedestrians. Hardware builders, software developers, licensor authorities, car owner, or the car itself? In fact, all of them should be accountable. Sensors are crucial for reactive behavior, software includes decision-making policies, authorities must be aware of the road conditions when accepting radically new driving modes, the car owner also accepts to be under those fully automatic conditions and the car itself could have been learning on its own how to change its own driving behavior.

Accountability goes hand in hand with responsibility. AI researchers and developers should take the responsibility to create models and algorithms to enable AI systems to reason about and take decisions in such a way that they can justify their decisions according to rational and logical principles. Current deep-learning mechanisms are unable to clearly explain why specific decisions come out of specific sets of inputs, and therefore we cannot really understand the rationale of the decision process. However, it may be useful that future self-driving cars have to deal with moral dilemmas as is posed, e.g., when the car is required to choose the lesser evil in a possible accident. Finally, Transparency means openness, a willingness to provide clear information about the designed algorithms as well as the need to describe and reproduce the mechanisms through which decisions are taken and adaptation becomes possible. It is evident that, if algorithms are not transparent enough when making relevant decisions on our behalf, we cannot judge where the responsibility lies and how can we argue against the quality of those decisions.

8 Conclusions

Stuart Russell, the well-known AI Professor at the University of California, Berkley, drafted and became the first signatory of an open letter calling for researchers to look beyond the goal of merely making artificial intelligence more powerful. "*We recommend expanded research aimed at ensuring that increasingly capable AI systems are robust and beneficial*'' [19]. Although some consider a myth that AI will either turn evil or conscious, we believe it is time to recognize the actual worry that AI is more and more turning competent, and simultaneously there is a possibility that its goals become misaligned with well-formed human goals.

Here, through this paper, and based on our limited experience, we intended to put forward some guidelines that could help specify and develop AI-based systems that go the right direction as well as try to demystify a few already current statements about unlimited possibilities of AI.

We remain excited about all the potential benefits of superintelligent either agents, systems, networks alone, or in cooperation with humans, and their respective relevant impact in the future human society. Meanwhile we believe that current glorification of AI is not proportional to the reality. That impact may still be decades away. We also adhere to Gödel's thesis in the sense that there will always be a truth that humans can comprehend but that a system will not be able to determine its truth or falsehood. Nevertheless, the scientific community in general and the AI community in particular, should be proud of launching all the interrogations that have to be made about the potential impact of AI in the future. The promoted symposium dedicated to the social and economic impacts of artificial intelligence in the next 10 years (AI Now), by the previous White House Administration, was a very relevant forum for discussing social, inequality, ethics, labor and health domains in which AI is raising pressing questions.

According to Kate Crawford and Meredith Whittaker [3], an uncomfortable truth has been revealed "*there are no agreed-upon methods to assess the human effects and longitudinal impacts of AI as it is applied across social systems. This knowledge gap is widening as the use of AI is proliferating, which heightens the risk of serious unintended consequences.*" It is also possible that spontaneous generation of synergistic control systems that will be no longer accessible to human control is nothing but another myth. But we should never forget that algorithms can be as biased as the data they draw on. As simple as that. Even if we look at the present, we are not willing to replicate what happened with Microsoft Corporation Chatterbot Tay that began to post offensive tweets, forcing Microsoft to shut down the service about 16 hours after its launch. In some specific scenarios, 16 hours could be too late.

In conclusion, I would like to emphasize that the main message the paper tries to convey is that it is at least smart, to start worrying about how to enforce human beneficial AI by using human intelligence to direct AI research in the benefit of humankind. We hope that, also in the future, ethical concerns will remain behind the law.

Acknowledgment

I deeply thank Professor Maximilian Etschmaier for his careful reviewing of this paper as well as for polishing my English.

References

[1] Beneficial AI Conference, "Asilomar AI Principles," https://futureoife.org/bai-2017/ .

[2] António Castro and Eugénio Oliveira, "The Rationale

Behind the Development of an Airline Operations Control Centre using Gaia Based Methodology," *Int. J. Agent-Oriented Software Engineering*, 2(3):350-377, 2008.

[3] Kate Crawford and Meredith Whittaker, Artificial Intelligence is Hard to See: Why We Urgently Need to Measure AI's Societal Impacts," https://medium.com/@katecrawford/arti_cialintelligence-is-hard-to-see-a71e74f386db, 2016, Online, Accessed 31-May-2017.

[4] Daniel Crevier, *Ai: The Tumultuous History of The Search for Artificial Intelligence*, Basic Books, 1993.

[5] Steve Crowe, "Robotic Trends," http://www.roboticstrends.com/site/author/Steve, 2017, Online, Accessed 30-May-2017.

[6] António Damásio, *Descartes' Error: Emotion, Reason and the Human Brain*, Avon Books, 1994.

[7] Subrata Das, "The Death of True Intelligence?", https://www.linkedin.com/pulse/deathtrue-intelligence-subrata-das/, 2017, Online; Accessed 11-September-2017.

[8] Virginia Dignum, Robots and AI are Going to Take Over Our Jobs! or Work with Us for a Better Future? https://www.linkedin.com/pulse/robots-ai-goingtake-over-our-jobs-work-us-better-future-dignum, 2017, Online; Accessed 30-May-2017.

[9] Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine will Remake our World*, Basic Books, 2015.

[10] Pedro Domingos, "Next 100 Years of Your Life," TEDx Talks La. https://vimeo.com/200120546, 2016, Online, Accessed 30-May-2017.

[11] EPFL, "The Blue Brain Project - A Swiss Brain Initiative," http://bluebrain.ep.ch/cms/lang/en/pid/59963, 2017, Online, Accessed 12-Septembery-2017.

[12] Edward A Feigenbaum, *A Personal View of Expert Systems: Looking Back and Looking Ahead*, Stanford University; Knowledge Systems Laboratory; No. KSL 92-41, 1992.

[13] Philipp Gerbert, Jan Justus, and Martin Hecker, "Competing in the Age of Artificial Intelligence," https://www.bcgperspectives.com/content/articles/strategy-technology-digital-competing-ageartificial-intelligence/, 2017, Online, Accessed 31-May-2017.

[14] Marcus Hutter, *Universal Artificial Intelligence*, Springer, 2005.

[15] Guideon Lewis-Kraus, *The Great A.I. Awakening*, The New York Times Magazine, Dec. 14, 2016,

[16] Eugénio Oliveira, "Robots as Responsible Agents," *Proceedings of The IEEE International Conference on Systems, Man and Cybernetics. Computational Cybernetics and Simulation*, 3:2275-2279, 1997.

[17] Arlindo Oliveira, *The Digital Mind: How Science is Redefining Humanity*, MIT Press, 2017.

[18] David Pereira, Eugénio Oliveira, and Nelma Moreira, *Formal Modelling of Emotions in bdi Agents*, Computational Logic in Multi-Agent Systems, Lecture Notes of Computer Science, 5056: 62-81. Springer, 2008.

[19] Stuart Russel, "Research Priorities for Robust and Beneficial Artificial Intelligence," Future of Life Institute, https://futureoife.org/ai-open-letter/, 2017, Online, Accessed 30-May-2017.

[20] John Searle, *Minds, Brains, and Programs*, Behavioral and Brain Sciences, Cambridge University Press, 1980.

[21] John Searle, "Watson Doesn't Know It Won on Jeopardy," https://www.wsj.com/articles/SB10001424052748703407305765415431312698767 4,2011, Online, Accessed 16-August-2017.

[22] Jean Tardy, *The Meca Sapiens Blueprint: A System Architecture to Build Conscious Machines*, Sysjet, Monterege, 2015.

[23] Stefan Wess, "Ai: Dream or Nightmare", TEDx Zurich Talks, 2014.

[24] F. Zambonelli, N. Jennings, and M. Wooldridge, "Developing Multi-Agent Systems: The Gaia Methodology", *ACM Transactions on Software Engineering and Methodology*, ACMr, 12(3):317 and 370, 2003.

Eugénio Oliveira is a Full Professor in Artificial Intelligence at the University of Porto and was Director of LIACC (Artificial Intelligence and Computer Science Laboratory) at the University of Porto until May 2016.

He got his PhD in Artificial Intelligence at New University of Lisbon in 1984. Awarded with Gulbenkian Prize for Science and Technology in 1984. "Guest Academic" at IBM/IEC in Belgium (84-85).

He supervised more than twenty PhD students in the area of AI and Software Agents. He published about 300 papers. H_índex (Google Scholar)=30, n. citations=4489 (4/12/17).

Current topics of interest include Software Agents architecture and strategies for cooperation, Trust and Reputation Models, Intelligent Transportation Systems, "Emotional-like" Agents, Text Mining and Multi-agent systems applications. He was General Co-Chair of the 18th EPIA Conference on Artificial Intelligent, September 2017.

People and Intelligent Machines in Decision Making

Kendall E. Nygard*, Md. Minhaz Chowdhury*, Ahmed Bugalwi*, and Pratap Kotala*
North Dakota State University, Fargo, North Dakota, USA

Abstract

The growing use of artificial intelligence methodologies in cyber-physical systems raises a host of issues in decision making. We examine cyber-physical systems in terms of levels of intelligent autonomy that they can possess, including special issues that arise in highly distributed and mobile systems and the advent of advanced machine learning. From the human side, we consider their competencies in conjunction with machines, including the errors and mistakes made by humans. Trust is a vital element from both the human and machine sides. We consider agent-oriented development paradigms and present models that capture the relationships between service levels and trust. Finally, we raise the modeling question of how to quantify reputation in relation to trust. We conclude that the advent of advanced machines that are enabled with artificial intelligence can bring great benefits, but there are also potential dangers.

Key Words: Artificial intelligence, intelligent machines, ethical machines, cyber-physical systems, trust, reputation, cyber security, autonomy, agents, deception.

1 Introduction

Modern society is engaged in a decision-making revolution driven by the rise of Artificial Intelligence (AI). Some AI decision applications now reach into areas that were very human-centric for a very long time, such as medical diagnosis, financial decisions, or conducting research into legal cases. Some are very personal and social, such as recommendations that turn into decisions for what movies to see, books to buy, or music to listen to. In the work that we report here, the focus is on cyber-physical systems, defined as those systems that integrate computers, networks, devices and physical processes. Cyber-physical systems are increasingly becoming AI-enabled. A few prominent examples are self-driving cars, drones, and industrial manufacturing equipment. These types of systems are heavily equipped with sensors, actuators, and controllers, but unlike their earlier generation counterparts, also involve integrated symbolic or sub-symbolic AI in their work. At an accelerated pace, humans are increasingly relinquishing some level of control of the machines and devices that serve them.

When people and organization arrive at mutually beneficial agreements and decisions, a high level of trust among them is required, even if the trust is formalized through a contract or other legal document. Similarly, if machines are heavily advising or actually making decisions, the people they serve must trust them. Even in the face of information, planning, and inputs that are ambiguous, uncertain, insecure, or imprecise, it is of critical importance that users have a basic trust in the AI-enabled systems that they are utilizing or with which they are interacting. Just as the industrial revolution was a profound and life-changing machine age with associated risks and skepticism, so the rise of AI marks a new machine age, again filled with risk, skepticism, and a need for trust. It is incumbent upon society to embrace such change and to understand and manage the risks, so that trust will naturally follow.

In this paper we first consider the elements of cyber-physical systems and the types of autonomy and decision-making that they can possess. We then turn to the elements of human-centered decision making, describe circumstances under which human error occurs, and identify types of human errors. This leads to approaches for assisting and augmenting human interactions with AI-enabled systems, but also leads to accompanying pressures to conform in certain ways. We then consider the software side, including systems that are operated by software agents; and relate these architectures and concepts to the functionalities of intelligent controllers. Finally, we more directly examine the concepts of trust, including possibilities of risks that can arise from deceptions and standardizations.

2 Cyber-physical Systems and Levels of Autonomy

A cyber-physical system (CPS) is a mechanism that integrates software components and networking with physical processes or devices. Some well-known examples include self-driving cars, drones, industrial manufacturing equipment, robot-assisted surgical devices, and weapons of war, smart phones, and the electrical grid. Cyber-physical systems are employed in all areas of importance to the well-being of humans and ecosystems of the world, including agriculture, energy, water, manufacturing, and health. In manufacturing specifically, the rise of CPS technologies accounts for numerous improvements in machine performance, including optimization of operations, remote control, prediction, and triggering of alerts for service needs, and remote diagnostics and self-healing.

Over the long history of machines and systems, their

* Department of Computer Science.

complexity and the tasks that they are charged with and are capable of performing has systematically advanced and increased. Systems referred to as autonomous are capable of carrying out certain functionalities on their own. For example, a typical industrial robot can be programmed to precisely carry out specific tasks involving such things as moving arms, gripping items, or laying down an almost flawless weld. Such a machine is clearly autonomous in the sense that it can act independently, but can hardly be deemed intelligent. In recent years advances in machine learning as a branch of AI provides ways for machines to be programmed to learn without the learning process itself being explicitly controlled. Machine learning methods accept, process and interpret data; and can identify patterns, classify situations, and make predictions. Driven by careful attention to configuration, a machine learning system that repeatedly ingests and processes more and more data is capable of automatically improving its performance, in effect accomplishing a reasonable definition of learning, producing a system that can be labeled as intelligent.

A truly intelligent system can systematically carry out decisions that result in goals being met without human intervention. A striking example is the Google DeepMind neural network that mimics the short-term memory of the human brain [16]. The accomplishments of the software suite provided by DeepMind includes a stick figure that begins with a cold start, then learns autonomously to walk like a human from point A to point B. Deep mind software can also learn to play video games, and to use machine learning to beat human professional players of the highly complex Asian game called Go. John Deere Corporation, the largest manufacturer of agricultural equipment in the world, recently invested $305 million in the development of a machine for farming crops like lettuce [17]. As the machine is driven in a field a software routine called Lettucebot accepts sensor inputs, identifies individual weeds and precisely directs a tiny spray of herbicide to the weed. Machine learning is employed to train the system to recognize the weed and even immature lettuce plants that will not be mature when the field is harvested. An important point is that the farming machine learns on its own. Already used in approximately 10% of lettuce farming, the technology accomplishes a 90% reduction in the use of herbicide. A variation of the machine for cotton farming is scheduled for release in 2018.

Autonomy can be described in terms of degrees or levels. For example, in automobile manufacturing, a number of manufacturers identify degrees of automation referred to as Advanced Driver Assistance Systems (ADAS). Although fully autonomous automobiles are not yet available, the National Highway Traffic Safety Administration (NHTSA) identifies levels of automation as shown in Table 1 [12].

A major issue in self-driving cars stems from the fact that control and responsibility are not the same thing. This explains the high interest in self-driving cars by the insurance industry. In the past, as levels of automation have advanced, there has been questioning, skepticism, and fear of the change. For example, even for a Level 1 technology, when anti-skid braking systems (ABS) were developed and later mandated, many people were apprehensive about a perceived loss of control. Today, there is no question that ABS technology, especially with digital controls that communicate through an on-board Controller Area Network (CAN), have saved many lives, although there are still a few specific types of accidents that could be averted if ABS were not in place.

3 Mobile Systems and Sensors

Some cyber-physical systems are inherently mobile, the prime example being smart phones. Although people may view themselves as being in control of their smart phones, in reality there is much autonomous functionality, such as location sensing using GPS chips, automatic updates, internet connectivity, photo management, invoking remote processes, and Bluetooth interconnectivity with other devices. Intelligent advisors that can learn from the behaviors of the user may also be on board. Considering Unmanned Air Systems (UAs), these aircraft must operate at a distance and out of physical contact from ground-based stations that control certain of their functions. Some are remotely piloted; others operate fully or partially autonomously. There is a strong trend toward operating of UASs with little human involvement or oversight.

Table 1: Levels of automation in automobiles

Level 0, No-Automation	The driver is in complete and sole control of all major vehicle controls at all times
Level 1, Function-Specific Automation	The driver is in overall control, but can cede authority to automation of one or more specific control functions (e.g., stability control, cruise control, automatic braking)
Level 2, Combined Function Automation	The driver shares control authority through automation of at least two primary control functions that operate in unison (e.g., adaptive cruise control and lane centering)
Level 3, Limited Self-Driving Automation	The driver cedes control of the vehicle to autonomous operation under certain conditions, but is available and expected to reengage driving when the vehicle provides an alert that indicates that there is a need to do so (e.g., coming upon a construction area)
Level 4, Full Self-Driving Automation	The driver provides input for a destination or routing, but is not expected to be needed or available for control at any time during the trip

The accompanying increase in the level of UAS trust is a basic challenge, raising the question of how to allow human controllers to override autonomous decisions if needed. In connection with cybersecurity for UAS, onboard decision-making software will have to be equipped to carry out intrusion detection from any source and preserve safety. Recent surveys show that there is a great deal of apprehension surrounding trust for self-driving cars [19]. Although much of the mistrust is centered on the potential for hackers taking control of the vehicle, some of the mistrust concerns the reliability and performance of the control systems. Privacy is also a concern.

A recent area of active research concerns the development of technologies for disabling, destroying or remotely taking control of UASs. Techniques include various kinds of radar detectors, audio and optical sensors, jammers, laser and electromagnetic pulses, and GPS spoofers. The technologies that are aimed at countering an autonomous aircraft also inspire techniques to defend against similar attacks, potentially increasing the trust level.

A great many cyber-physical systems rely upon sensors that collect data on system health and autonomously make decisions based on the sensed data. For example, in the Smart Grid, data from Phasor Measurement Units (PMUs) and other sensors can form the basis for autonomous self-healing actions that reroute power or trip critical breakers in the face of device malfunctions or terrorist attack, preventing cascading failures. Since such failures propagate at the speed of light, human operators in control centers are very limited in their ability to take appropriate actions quickly enough to prevent disastrous ripple effects. So, this is an area where machine intelligence is rapidly overtaking human intelligence. These issues also apply to many other types of cyber-physical systems, including self-driving cars, UASs, sensor nets, and multi-purpose robots. The trustworthiness of the sensor data in such applications is dependent upon characteristics of the data, the sources, communication paths, and sensor fusion algorithms. However, it can be difficult to differentiate between problems that are caused by failures of devices, sensors, or the communication network and problems due to malware or intrusions initiated by hackers.

Supervisory Control and Data Acquisition (SCADA) systems are time-honored legacy technology for controlling critical infrastructure that almost all people in the United States rely upon every day for provisioning basic needs like water, power, heat, and waste disposal. As these control systems are reconfigured for remote access and internet connectivity, they become more exposed to attackers and compromise, which diminishes trust and increases risk. Since many of these systems were designed under an assumption of air gap isolation, with no expectation of internet connectivity, there are big issues concerning the interaction of humans with SCADA systems to improve trust and defend the cyber-physical infrastructure.

4 Human Decision Making and Errors

People are impressed by heroic acts, such as when Captain Sullenberger took control of a highly instrumented Airbus A320 that was disabled by bird strikes and successfully landed the aircraft on the Hudson river [11]. The success of the landing is a tribute to the ability of a person to stay calm under duress, receive and process information about a difficult situation, determine what to do in response, and carry out appropriate action. Education and training can enhance the ability of a person to perform well under duress. However, human performance is often fraught with mistakes. As far back as 1980, starting with the work of Reason [14], followed by work of Rasmussen [13], ways in which humans receive and process information and arrive at decisions in complex situations were analyzed. Figure 1 illustrates a high-level breakdown of human activity within complex environments

When people commit errors or exhibit low performance there are usually human-centric explanations. For example, for an error that is traceable to an issue in sensing and perception on the part of a person, the explanation might be fatigue or an interruption in attention. In processing and decision-making, an explanation might be lack of training or knowledge or a memory lapse. In taking action, an explanation might be a shortfall in a skill or a strict adherence to a well-learned routine that does not work for every circumstance. A study of Smith [20] showed that while operating machinery or industrial equipment, a trained person who is carrying out a non-standard task will create an error condition by failing to heed an adverse warning indicator 10% of the time. Such potential for human error has led to the development of many types of ideas and systems for reducing the occurrence of mistakes, especially in situations where safety is a key concern.

5 Human Competencies, Autonomy, and Machine Learning

In the context of humans carrying out tasks, there is a great deal of attention paid to identifying competencies associated with different types of jobs. As applied to the use of cyber-physical systems, this includes such things as specific kinds of training, adherence to protocols, and accuracy of work processes.

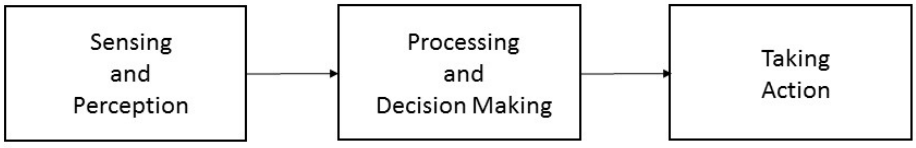

Figure 1: Stages of human activity in complex environments

Behavior Marker systems [10] provide a structured approach to reducing human error. Behavior Markers are defined as "observable, non-technical behaviors that contribute to superior or substandard performance within a work environment." In an application domain, they are derived by analyzing data regarding performance that contributes to successful and unsuccessful outcomes. Such a system provides an observation-based method to capture and assess performance based on data rather than on gut feelings. The approach is highly structured, is based on behavior scoring methods, and serve as the basis for creating checklists that ensure that important steps and procedures are properly followed. Behavior Marker systems have found success in the work of surgery teams and aircraft safety teams preparing for takeoff, anesthesiologists, and software developers. By imposing structure into important and complex multi-step working environment, Behavioral Marker systems can help to produce consistently good outcomes and avoid bad outcomes that arise from people taking shortcuts, overlooking steps, or handling tasks poorly. However, Behavior Marker systems may not work well when people are forced into rapid decision-making or must deal with unforeseen events.

6 Human-on-the-Loop

A Human-on-the-loop in a semi-autonomous cyber-physical system may be responsible for setting parameters or intelligent objectives, monitoring sensor settings continuously or intermittently, or possibly to step in to control loop operation at times.

Figure 2 illustrates the role of multiple sensors in monitoring a semi-autonomous Cyber-Physical system that is continuously evaluated for system health and governed during normal operation by a controller that can operate autonomously and adapt to changing conditions. The Figure also incorporates the influence on control that a human operator can carry out by receiving information and alerts and taking action to parameterize, set goals, or otherwise influence the controller. This type of Human-on-the-loop is characterized by an intermittent supervisory control such as is implemented in systems like air traffic control, modern fighter aircrafts, crisis response, or some process controls in manufacturing. The connection with systems engineering is apparent, as the human is linked via the implementation of the controller that in turn governs the action of actuators that actually make changes in the operation of the cyber-physical system. We note that the information that the human receives may be detailed readouts or be highly visual or audible alerts.

Another point is that any of the devices is subject to component failure or a hacking attack, which is also illustrated in the figure. In addition to sensor attacks that change or otherwise interfere with sensor measurements, there can be attacks on actuators, any of the message passing actions, or the controller itself. One issue is that at the sensor level, most sensor fusion procedures cannot distinguish attacks from faults. For example, an out of range sensor measurement could be transient and explained by noise, or, if persistent, could be either a failing device or an attack. Non-transient faults that persist and manifest themselves as consistent readings either too high or too low are often an attack, though not always. In addition to disruptions to sensor readings that show abnormal out of range

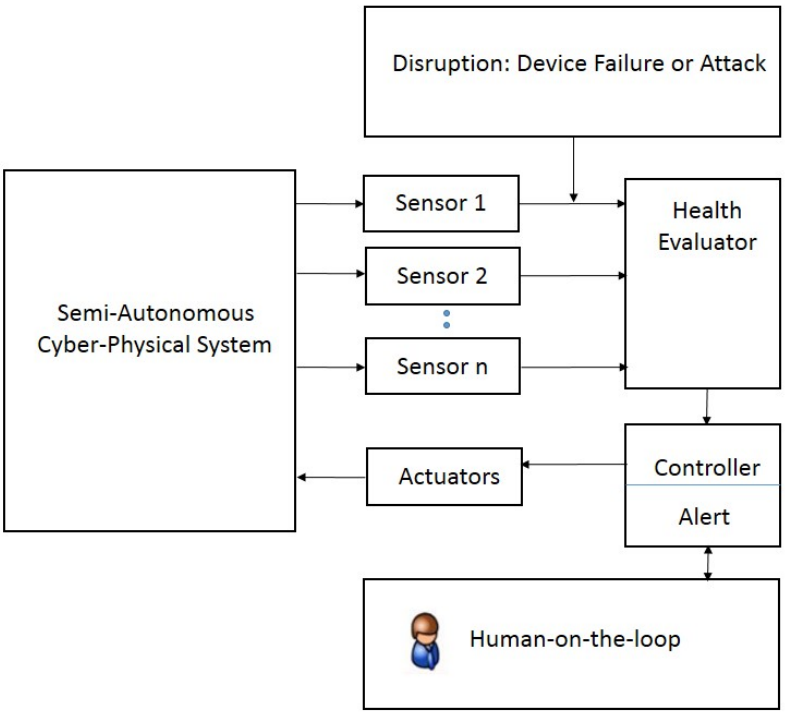

Figure 2: Human on-the-loop

readings, there is also the possibility that disruptions could report reading that are normal when in fact they are not. In systems like the smart grid, intelligent and nearly instant controlling capabilities are important to avoid disasters like cascading failures, since human-on-the-loop actions in such applications would find it nearly impossible to act quickly enough to trigger self-healing controller actions, such as rerouting power or strategically throwing breakers. In such applications, autonomously operating AI-enabled controllers have great potential to outperform people.

7 Defining and Measuring Trust

Grandison and Sloman define trust as the firm belief in the competence of an entity to act dependably, reliably and securely within a specific context [7]. In this sense, if a person has trust in an entity that is another person or a specific machine or system, there is confidence that the entity will deliver performance as expected. Trust, then, can be a noun that captures a belief. However, trust can also be used as transitive verb, which in this context means that there is an object (person, machine or system) that is to be trusted. This could be written in the form A → B, to carry the meaning that A does trust B to fulfill some purpose. This also implies the trust is contextual, specific to some domain with intended goals and purpose. The purpose may be to access resources or information, control or monitor a process, provide some service, or make a decision. In online systems, trusted message passing is a phrase often used to describe public/private key encryption, including digital signatures. However, this restricts trust to the meaning that the message got through from sender to receiver with no issues of interception, modification, etc. Good cybersecurity is important in helping to ensure trust, but unacceptable outcomes can and often do occur even when all of the communication between A and B is accurate and fully secure. These unacceptable outcomes destroy trust.

In formalizing trust, discrete levels can be defined for an otherwise continuous variable over a range such as (0.00, 1.00), which basically follows the fuzzy logic concept of soft computing. Accordingly, we can then assign verbal descriptions to values of trust, as illustrated for example in Table 2.

Table 2: Fuzzy levels of trust

Value	Description
(0.90, 1.00)	Very High Trust
(0.75 , 0.90]	High Trust
(0.50 , 0.75]	High Medium Trust
(0.25 , 0.50]	Low Medium Trust
(0.10 , 0.25]	Low Trust
(0.00 , 0.10]	Very Low Trust

Trust metrics can be functions derived from service or task performance levels, accounting for trust to increase when outcomes are consistently good, and decrease when trust is repeatedly bad. As applied to cyber-physical systems, it is critical that AI-enabled machines be highly trusted by those that use and operate them.

Trust is often built on evidence, which can then intertwine trustworthiness with reputation, which is elaborated upon later in this paper and also dealt with in work reported in [1,23]. Trust is also inherently a broader concept than system security. This is fundamentally because a compromised system, even if it appears to be usable, cannot possibly be trusted. Compromise can take many forms, including system slowdown, incorrect outputs, system shutdown, theft of information, etc. Severe security vulnerabilities exist in any device connected to the internet, and are especially difficult to detect and defend against in highly distributed networked systems that include wireless connectivity. Consider, for example, the attack of October 21, 2016, in which the Mirai botnet brought down multiple sites, including Twitter, Netflix, Reddit, and CNN and many others throughout the United States and in Europe [22]. The technique used was distributed denial of service (DDoS). Basically, the botnet invaded Internet of Things (IoT) devices, such as digital cameras, DVD players, smart TV sets, smart electric meters, smart phones, process sensors in power plants, and commercial security cameras. Mirai did its malicious work by invading around 100,000 IoT devices, which in turn launched coordinated message traffic that swamped servers run by Dyn, a company that ran much of the internet domain name system (DNS). Concerning the scale of the Internet of Things, Estimates vary greatly, but Gartner estimates that in 2017 the number connected devices is around 8.4 billion and will reach 20.4 billion by 2020 [6]. With so many IoT endpoints to choose from to invade and compromise and force to do their bidding, the scale of this type of attack is truly massive. Since most cyber-physical systems are connected to the internet and many have wireless connectivity, there is great vulnerability to compromise, with the accompanying loss of trust.

The concept of machine resilience in design and operation is also intertwined with trust. For example, what happens if a hardware or software component of a machine is compromised, fails, or is incorrectly instantiated? A highly resilient machine will not cause a disaster, fail gracefully, self-heal, or continue to provide required service by some means. High resilience may be the result of excellent machine design by a person, or, alternatively, the result of excellent intelligence on the part of the machine. In situations that are inherently people trusting machines, it can be asserted trustworthiness on the part of the machine within a context is a binary outcome – either the machine is trustworthy within a domain or it is not [23].

8 Machines Trusting People

We have asserted that a user of a system, from the viewpoint of cybersecurity and more, must have at least a reasonably high level of trust in the system. On the other side, and anthropomorphizing, we also assert that the system itself must have some means of trusting the user. The traditional method is simply access control. Authentication factors that can verify a legitimate user fall into three categories: 1) Knowledge factors that are based something that the user should know, like a valid

and current login and password, 2) Possession factors that include things that the person has with them, such as a key fob or a scannable ID card, and 3) Inherence factors that are specific to an individual, such include biometric traits like a retina or fingerprint scan. Two and Three factor authentication where more than one factor is used is now common, and greatly improves access control. However, more broadly, an autonomous and intelligent machine that gets an instruction and control from a human user may require a form of authentication that goes beyond the usual verifications steps. It may even be the case that the machine would have choices as to which human it should empower to complex their side of a task within a domain, making the "machine trusting man" decisions quite complex.

For an intelligent machine to truly be confident in empowering a user, it may be necessary for the machine to have some ability to determine if the actual behaviors of the user are legitimate and appropriate, even if the user has been through a standard authentication screening and granted access. Research into the design and use of the mechanisms by which users control and interact with systems comes into play. For example, consider a user who is monitoring, interfacing, and has some control and interaction with an industrial process machine. Working from a computer console and employing a voice recognition system, the person has responsibility for carrying out various operations by providing mouse movements, clicks, keystrokes, and audible inputs. There may also be buttons that can be pushed, levers that can be pulled, etc. People vary considerably in attributes such as the time taken between clicks, path of cursor motion when the mouse is moved, hovering time over icons, the nature of their spoken voice, etc. Different cognitive processes translate into differences in how individual people interact with the machine, to the extent that usage patterns can be as distinctive to an individual as a biometric like their fingerprints [4]. An intelligent machine can potentially enhance safety and security by recognizing and responding only to people who are authorized to work with the machine interface and can also verify that the user is not impaired by things like fatigue, illness, confusion or any factor that potentially triggers human error. Imposters and hackers can and should also be recognized and be invalidated. In designing the interface with an intelligent machine, there are many principles that should be followed, such as the machine providing information feedback that confirms every action, prompts the steps of sequences, provides error handling, the means to reverse actions, and ensures that memory load imposed on the individual are within bounds. There may also be differences in the interfaces and the empowerment available among users in different authority domains. In short, in the era of intelligent machines, it will be important for the operator to trust the machine, but also for the machine to trust the operator. But modeling human behavior is a daunting task, and expecting machine intelligence to distinguish between legitimate and unauthorized operators or even adversaries involves more than a simple checklist.

9 Agents

An intelligent software agent is an encapsulated software system situated in an environment where it can conduct flexible and autonomous actions to meet its design objectives [9]. Agent goals may be shared or private. The key characteristic of a software agent that is distinctive from other programming paradigms is that agents persistently assimilate and analyze information, evaluate the options available to them, choose among them and act. In software, this precisely mirrors the human decision-making process described above. In contrast with agent systems, other programming paradigms are much more prescribed. In an intelligent agent architecture, the agents carry out roles such as monitoring the activities and intentions of people and other agents and devices, brokering the completion of tasks utilizing available resources, following preferences that have been programmed in or received externally from people, or agreed upon through a negotiation. Intelligent machines have agents that are heterogeneous and cooperative. Heterogeneity comes from devices sensing a variety of different things, activating diverse controllers, negotiating the completion of various tasks, and finding acceptable solutions to various models. For example, in the self-driving car example, lane centering and cruise control must cooperatively work together to accomplish their prescribed tasks.

In the belief-desire-intention (BDI) agent model, beliefs correspond to the state of the agent, desires to the effects that the agent attempts to cause and the intentions of the plans or steps that the agent is following [8]. Figure 3 illustrates a generic multi-agent system in which some agents stand alone and others are composites. Some agents may be mobile, which is also illustrated in the figure.

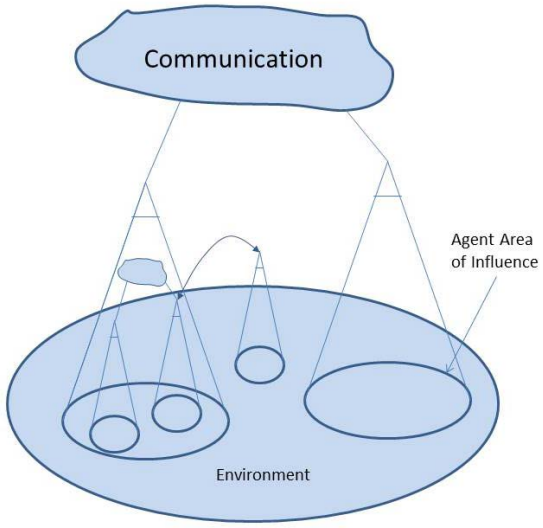

Figure 3: Generic multi-agent system

Some agents can be charged with being information couriers, responding to the preferences and interests of people and other agents by providing information upon request or subscription. Unlike most software developed today, agent software is proactive and can run autonomously and continuously, potentially relieving the human from various tasks. Fully self-driving cars will require almost nothing from the operator/passenger, with software agency providing the transportation service. But in the spirit of the human-on-the-loop, if the passenger decides to change their destination enroute, there must be a provision for a software agent that is tracking and is responsive to the person who inputs a new instruction. The terminology "agent" is inspired to emphasize the fact that agent software can act on behalf of a person or other agent, providing relief to that other party. Note that if a person is to be truly served by an intelligent agent, the person must understand what the agent can carry out, and also trust that the agent will do so. Only then will the person feel that they are the one actually in control. This raises the question of responsibility. If we insist that intelligent machines bear the responsibility for dangerous or failure conditions or receive the credit for good outcomes, then we raise the question of whether machines can exhibit ethical behaviors.

Intelligent agents with goals and beliefs require cognitive capabilities. Dillenborg [5] proposes the Socially Distributed Cognition (SDC) metaphor which essentially asserts that a human-computer pair involved in shared problem solving should function as a single cognitive system. Following this metaphor leads to the concept of a cooperative human/machine society agent society in which all parties interact and participate to achieve a shared goal. If an intelligent software agent on the machine is equipped with machine learning models or data mining techniques, and natural language processing, then that agent can gain knowledge continuously from various sources. The agent may seek to identify patterns, and over time can evolve and develop behaviors and responses based on the newly acquired knowledge. In this way, agents can learn and anticipate new problems that lead to new counter solutions. For example, automobiles are already equipped with specialized sensing devices designed to recognize attacks and counteract them. A shortcoming today is that the patterns that the devices look for are pre-programmed at the design stage. However, if the automobile is instead equipped with machine learning procedures, the cognitive agent can potentially recognize new threats and learn how to counteract them. Basically, the cognitive agent would be equipped to receive streams of scenarios and situations from various sources, and use them to shape and grow its knowledge base to include the new threat patterns, followed by developing the strategies to counteract them. Accomplishing both descriptive and predictive capabilities is a tall order. But may well be the Holy Grail for developing threat defenses, especially for self-driving cars. A high trust relationship is essential when a human or agent delegates tasks to intelligent agents because there is uncertainty about how a cognitive agent will behave and evolve over the long term. For instance, is there a way to guarantee that a software agent, after being exposed to multiple malicious actions, will not treat them as normal after a time, and essentially turn on its human master? Consider a hypothetical military operation in which an agent equipped with software like that of Google DeepMind has learned how to guide a missile to a target in the same way that it can guide a stick figure learning to walk. Could there be a guarantee that the missile will have learned the appropriate parameters to achieve the intended outcome? Could an intruder that is possibly malicious or a design flaw have fed the machine learning missile some false scenarios that lead the fired missile to change the assigned target and bomb the wrong thing?

It is evident that trust among operators and agents is a vital element that must be supported in any human/machine environment in which there are risks and in which malicious attacks can occur. This argues for mechanisms that in some way ensure that the agents are trusted and will act on behalf of the trustor in carrying out tasks. Also, if the trusted agent receives sensitive and private data, the trust models in such systems should provide mechanisms to give operators confidence that they are dealing with highly trustworthy agents and their data is protected and secure.

10 Trust, Deception, and Reputation

A 2016 survey sponsored by Google for the Economist Intelligence Unit indicates that 99% of 552 respondents in a survey do use cloud services, but only 16% report that their trust in the cloud is very high within their organization. [21]. The basic reasons for mistrust are uncertainty about the following items: 1) where their valuable data reside, 2) who can currently see the data, 3) who has seen the data in the past, 4) whether the data is untampered, 5) where the processing is performed, 6) how the processing is carried out 7) how backups are done, and 8) whether there will be issues in access such as delays. In this study, those respondents who had higher trust in their cloud provider also showed significantly higher profits. This correlation, however, does not necessarily mean that trust and profit go hand in hand. The correlation may simply mean that highly trusted cloud providers tend to be those that do provide excellent service and are thus helpful in generating profits. But there could easily be cloud providers for which trust is both high and misplaced, but manages through other supporting activities (which may even be unethical) to generate high profit.

There are two approaches to quantifying trust, particularly when there are possibilities for deception on the part of one or both parties. In a cognitive approach, belief states are the foundation of trust. As described earlier, fuzzy logic can be used to provide a quasi-quantitative basis for measuring trust. Deception influences placement on such a scale. In a competing approach, trust can be quantified by contrasting expected versus actual action on the part of one agent in interaction with another. A wide variety of methods for quantifying trust can be developed, including overlaying a game theoretic approach in which a utility function that aggregates the results of past interactions between two agents is developed [15, 23].

In cloud computing, trust can be measured in terms of lapses of the level of service provided, as measured by a specific

service attribute. Typical example service attributes include availability, dependability, security, usability, accessibility, computational performance and reliability. Lapse of these service attributes are measured from their expected value from the perspective of the consumer. These values are quantitatively specified in a Service Level Agreement (SLA). Although cloud services are not a cyber-physical system, they do provide a natural means of structuring a model for systems in which a service is provided, expectations are in place, and there are intuitive ways to quantify relationships between users and providers.

Figure 4 illustrates a level of service metric for a particular service attribute at discrete points of time t0, t1… t10 and the corresponding trust value, as an example of trust and a service metric value relationship. The service metric value is normalized into the range [0, 1] and the trust value falls into the range [-1, +1]. The trust value is calculated by comparing actual versus contracted values of the service metric as defined in an SLA. The expected service metric metric of this example is 0.5. The figure reveals that at certain points of time, for example at t2, t3, t7, t8 and t11 the service falls short of the contracted value, which can be viewed as a deception. Accordingly, the corresponding trust values fall in response. Similarly, whenever the service metric value is above the threshold of 0.5 the trust value goes up in response. This happens at points in time t4, t5, t6, t9, and t10.

Each specific service attribute has its own metric. For example, response time is the metric for availability of a cloud service, Millions of Instructions per Second (MIPS) is the metric for computational performance provided by the Virtual Machine on the cloud. We have conducted experiments with these two-service metrics using a cloud simulator [2]. When lapses occur, there is a need to model increments and decrements in trust that occur, including repetitive strings of service that meets standards or of lapses that do not. Strings of good service that lull the user into becoming satisfied and complacent but followed by periods of bad service are referred to as a con man trick. In essence, the provider seeks to maximize profit by minimizing their operating cost. This type of behavior can occur in cycles. This fits with the game theoretic approach to quantifying deception.

An old Chinese idiom can be stated as "Fool me once, shame on you, and fool me twice, shame on me." We follow a variation the lines of "Fool me repeatedly, my trust in you diminishes. Satisfy me repeatedly, my trust in you increases." Evidence suggests that trust decrements should be more aggressive than increments since noticeable repeated bad service tends to rapidly become troubling to users, and repeated good service tends to be less noticed. Hence, we experimented with various algorithms that increments trust more slowly than it decrements. Each algorithm follows a mathematical model that updates a trust value after each interaction of the consumer with the service provider. One parameter in the model is sensitivity to defection (inadequate service), and another is sensitive to cooperation. When defection occurs, both of these parameters are adjusted by amounts that depends upon the previous value of these parameters individually, the previous trust value and a constant. The cooperation parameter is represented by α and defection parameter by β. When cooperation occurs and the previous trust value is positive, then the trust value is incremented at rate α. When defection occurs, the trust value is decremented by β. Initial values of α and β are at a specified ratio [3]. These initial values and their ratio form the basis for hypothesizing that con man behavior is occurring, but also suggests that con man behavior tends to reach limits at which the deception loses its effectiveness, in that the "victim" is no longer easily fooled. Realistic mathematical schemes for adjusting α and β show that when a cycle of defection and cooperation perpetuates, the trust value systematically goes progressively lower on the cycles. Recovery of the trust value into the positive range can require a long period of time with cooperative interactions, so the potential for the provider to gain from con man behavior tend to progressively become more and more difficult. Figure 5 illustrates this type of cumulative effect of repetitive defection across cycles [3]. The height of service metric glitches represent the lapse amount of service. This illustrates how important trust is in human/machine interaction.

We have used variations of this algorithm in simulation of

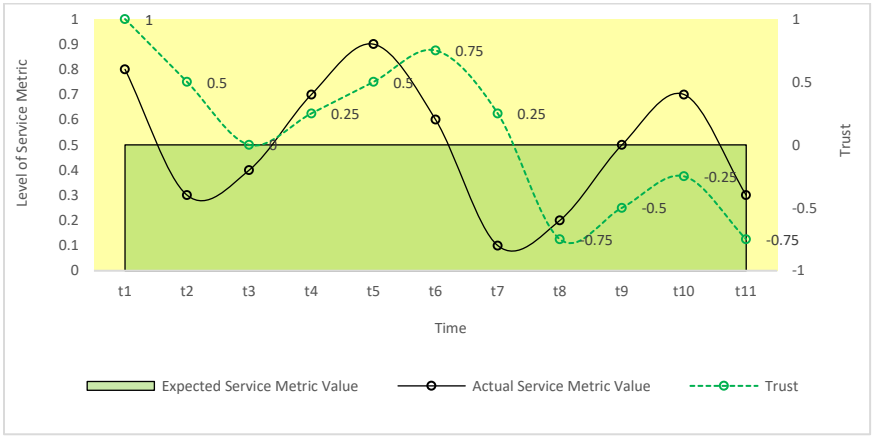

Figure 4: Example relationships between service and trust

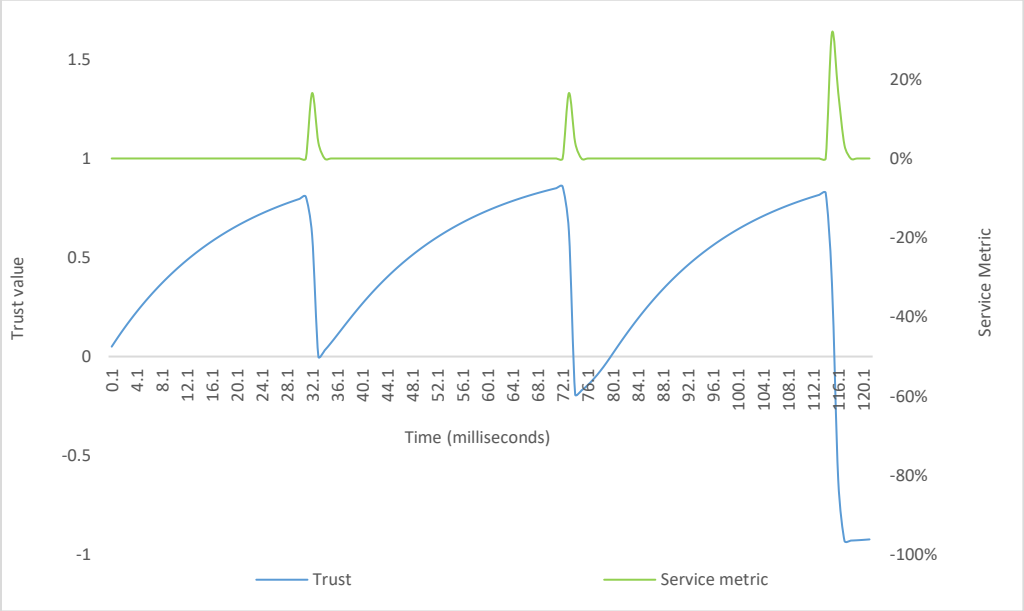

Figure 5: Repetitive defections induce unrecoverable low trust

cloud computing domains to monitor and identify con man deception in cloud services for both Software as a service (SAAS) and infrastructure as a service (IAAS). In both cases, the con man deception can come from the service provider, which is likely an autonomous agent, but with perhaps some human control.

In related work we are also experimenting with direct models of reputation, which is closely related to trust. Here one component that we consider is the context of agent interactions, a measure of the relative importance of the interaction. This can be placed on a fuzzy scale, ranging from very low to very high importance and normalized to [0,1]. A second component is the age of the interaction, with recent occurrences scored higher than older ones. A third component is a measure of satisfaction, which can range from dissatisfied to highly satisfied. Combing these factors yields the following expression:

$$T(g) = \frac{\sum_{i=0}^{n} C_{context}(i) * C_{age}(i) * C_{sat}(i)}{n}$$

Where

g = The targeted agent
T(g)= The trust value of the target agent
n = The total number of past communications
$C_{context}(i)$ = The communication context value
$C_{age}(i)$ = The age of the communication
$C_{sat}(i)$= The satisfaction of the communication

This approach to reputation can provide an ingredient into devising a trust value applies in certain situations in which repetitive interactions occur. It is inspired by reputation systems that are in common use for scoring retail buying transactions on the web, but mapped more specifically to human/machine systems.

11 Ethical Machines

Finally, we consider the question of ethical machines. If machines are to make decisions and intelligently carry out tasks, they must be trusted like reliable and ethical humans. Utilitarian ethics [18] is an approach to making ethical choices by balancing positive and negative utility outcomes. In the case of self-driving cars, utility is in terms of human welfare. Advocates for the development and deployment of cars that reach Level 4 automation strongly argue that safety will be dramatically improved, thereby on balance increasing human welfare. This is tantamount to arguing that problematic behaviors on the part of humans behind the wheel cause many accidents that can be eliminated. Since people, at least in theory, act in accordance with moral and ethical principles, this leads to the issue of whether machines can be design and develop to act in accordance with established ethical standards. However, in many cases humans cannot agree on how to evaluate and score positive and negative outcomes, making it much more difficult to expect machines to be ethical. This inevitably leads to the conclusion that mankind has little choice but to have a role in human/machine decision making, even if the machine is deemed to be highly intelligent.

12 Conclusion

Our work considers a range of issues in human/machine cyber-physical systems. We recognize that AI-enabled machines are becoming much more common, particularly with capabilities of sub-symbolic machine learning that allows them to learn and perform from experience. We consider examples

like self-driving cars as a prototypical cyber-physical system for which there is rapid development. Human-on-the-loop systems are common today, and represent control systems that are largely in operation but provide for human activity and intervention under prescribed conditions. In recognition of human limitations and the propensity of people to make mistakes and cause accidents, we explore human-centric decision processes and compare them to the work of AI-enabled machines. Trust is important in human/machine systems, but trust is an amorphous concept that is difficult to pin down and quantify. Trust has a connection to cybersecurity, but encompasses much more. Trust can also be turned on its head to consider how machines might trust people in addition to how people trust machines. Agent architectures provide an appropriate way to program cyber-physical systems to capture the dynamics of how people and machines can work together. We also consider deception and reputation in relation to trust, including the potential for machine learning agents to learn the wrong things, be misled, or carry out deceiving behaviors. We see a great need for theoretical work aimed at ensuring that AI systems produce quantifiable results that lie within well-understood or prescribed bounds. A few results along these lines are available for specific types of methodologies, including genetic search and certain emergent intelligence approaches. But much more work along these lines is needed before humans will place high levels of trust in AI-enabled machines. Overall, we conclude that there is considerable risk associated with AI-enabled machines making decisions and exercising control, particularly in dangerous settings. More specifically, in types of artificial intelligence like machine learning, there remains potential for learned behaviors that produce outcomes that are not only unforeseen, but may be impossible to guarantee are even within normal bounds. In downside risks, even if probabilities of the bad outcomes are small, they can potentially be truly disastrous. An example of a catastrophic outcome is if a missile in a war scenario that has learned to recognize and destroy enemy armored personnel carriers by mistake strikes a school bus with innocent children aboard. Accordingly, we recommend caution and great care in developing such systems, to mitigate the risk that unforeseen and potentially harmful outcomes could occur. Nevertheless, the rise of intelligent machines is very real, has great potential for good, and has a powerful momentum.

References

[1] C. Castelfranchi and Y-H Tan, "The Role of Trust and Deception in Virtual Societies," *International Journal of Electronic Commerce*, 6(3):55-70, 2002.

[2] M. Chowdhury and K. Nygard, "An Empirical Study on a Con Resistant Trust Algorithm for Cyberspace," *Proceedings of the 18th International Conference on Internet Computing and Internet of Things*, pp. 32-37**,** July 2017.

[3] M. Chowdhury and K. Nygard, "Deception in Cyberspace: An Empirical Study on a Con Man Attack," *Proceedings of the 16th Annual IEEE International Conference on Electro Information Technology*, pp. 410-415, May 2017.

[4] Y. Deng and Y. Yu, "Keystroke Dynamics User Authentication Based on Gaussian Mixture Model and Deep Belief Nets," *ISRN Signal Processing*, 2013:1-7, 2013.

[5] P. Dillenbourg, *Distributing Cognition over Humans and Machines*, S. Vosniadou, E. De Corte, B. Glaser and H. Mandl (Eds.), "International Perspectives on the Psychological Foundations of Technology-Based Learning Environments," Lawrence Erlbaum, Mahwah, NJ, pp. 165-184, 1996.

[6] Gartner Inc. Report, "8.4 Billion Connected 'Things' Will Be in Use in 2017," https://lp.google-mkto.com/rs/248-TPC-286/images/EIU_Trust_in_Cloud_Technology_FINAL.pdf, February, 2017.

[7] T. Grandison and M. Sloman, "A Survey of Trust in Internet Applications," *IEEE Communications Surveys and Tutorials*, 3(4):2-16, October 2000.

[8] A. Guerra-Hernández, A., Fallah-Seghrouchni and H. Soldano, "Learning in BDI Multi-Agent Systems," International Workshop on Computational Logic in Multi-Agent Systems, 2004.

[9] M. Hamdi, S. Alqithami, and H. Hexmoor, *On the Core of Agents Interactions in a Spontaneous Networked Organization*, Elsevier SerVerse Science Direct, 2013.

[10] L. Lacher, G. Walia, F. Fagerholm, M. Pagels, K. Nygard, and J. Münch, "A Behavior Marker Tool for Measuring Non-Technical Skills of Software Professionals: An Empirical Study," *International Journal of Software Engineering and Knowledge Engineering*, 25(9-10):1733-1738, November 2015.

[11] B. Mutzabaugh, *As 'Sully' Debuts, a Look Back at the 'Miracle on the Hudson'*, USA Today, September 9, 2016.

[12] *National Highway Traffic Safety Administration*, U. S. Department of Transportation Report, Automated Vehicles for Safety, September, 2017.

[13] J. Rasmussen, "Risk Management in a Dynamic Society: A modelling problem," *Safety Science*, 27(2-3):183-213, 1997.

[14] J. Reason, J., *Human Error*, Cambridge University Press, New York, USA, 1990.

[15] A. Salehi-Abari and T. White, "Towards Con-Resistant Trust Models for Distributed Agent Systems," *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, 272-277, 2009.

[16] T. Simonite, *How Google Plans to Solve Artificial Intelligence*, MIT Technology Review, March 2016.

[17] T. Simonite, *Why John Deere Just Spent $305 Million on a Lettuce –Farming Robot*, Wired, September 2017.

[18] P. Singer, *Practical Ethics*, 2nd Ed, Cambridge University Press, 1993.

[19] M. Sivak and B. Schoettle, "Cybersecurity Concerns with Self-Driving and Conventional Vehicles," University of Michigan Sustainable Worldwide Transportation Report SWT-2017-3, February 2017.

[20] D. Smith, *Reliability and Maintainability and Risk*, 7th Edition, Elsevier, 2015.
[21] C. Whelan, "Trust in Cloud Technology and Business Performance: Reaping Benefits from the Cloud," Report of The Economist Intelligence Unit, pp. 1-27, June 23, 2016.
[22] N. Woolf, "DDoS Attack that Disrupted Internet was Largest of its Kind in History, Experts Say," The Guardian, https://www.theguardian.com/technology/2016/oct/26/ddos-attack-dyn-mirai-botnet, 2016.
[23] J. Urbano, A. P. Rocha, and E. Oliveira, *A Socio-Cognitive Perspective of Trust*, S. Ossowski, Editor, Agreement Technologies, Law, Governance and Technology Series, 8:419-431, Springer, Dordrecht, 2013.

Kendall E. Nygard is a full professor and serves as department Chair of Computer Science at North Dakota State University (NDSU). He is also the founder and Director of the NDSU Institute for Cybersecurity Education and Research. He has served in Washington D. C. as a Jefferson Science Fellow at the U. S. Department of State and Senior Science Advisor at USAID. He is currently a U. S. Department of State Virtual Fellow. Dr. Nygard is also a fellow of the International Academy, Research, and Industry Association. He is a recipient of the 2016 Chamber of Commerce NDSU Distinguished Faculty Service Award. Dr. Nygard earned his PhD degree at Virginia Polytechnic Institute. He has advised 23 PhD students and more than 150 Master of Science students. His research is widely published. Application areas in which he has conducted research include cyber security, smart electrical grid, sensor networks, unmanned air systems routing and scheduling, wireless ad hoc networks, encryption, and social media. His primary methodologies are big data analytics, optimization models, artificial intelligence, and simulation.

Md. Minhaz Chowdhury is a Ph. D. candidate in the Computer Science Department at North Dakota State University. His research includes trust, deception, cyber security, smart electrical grid, and analytics concerning perceptions of genetically modified organisms. He holds a BS in Computer Science and Engineering from Daffodil International University and received the Vice Chancellor's Gold medal award from the President of The Peoples' Republic of Bangladesh. He received his MS in Computer Science from North Dakota State University and has worked as a programmer at Blue Cross Blue Shield of North Dakota. His PhD dissertation research is in cyber security with the focus being computational trust as a tool to identify deception in cyberspace.

Ahmed Bugalwi received the Bachelor of Science degree in Computer Science from the Benghazi Institute of Technology and the MS degree in Computer Science from The Libyan Academy. Currently he is a PhD candidate in Software Engineering at North Dakota State University. Ahmed's background includes approximately 10 years in the software industry. Before joining NDSU, he worked as a lead developer and the head of technical support department at the Libyan Stock Market.

Pratap Kotala is a Senior Lecturer in Computer Science at North Dakota State University. In over 18 years he has taught a wide variety of Computer Science courses at both undergraduate and graduate levels. He has two years of experience as an IT consultant in the Health Insurance industry. His research interests include Cybersecurity, Software Engineering and Data Mining. Dr. Kotala serves as workshop leader for summer camp programs in cyber security, including special summer camps for middle and high school girls and Native Americans. He has served as organizer and panelist for North Dakota Cyber Security professional and academic conferences. His is a participant in the Stanford University Hacking for Defense program on lean methodologies. Dr. Kotala earned his Ph.D.in Computer Science from NDSU.

Index

Authors

A

B

C

D-G

H-J

K-L

M-O

Key Words

Instructions for Authors

The International Journal of Computers and Their Applications is published multiple times a year with the purpose of providing a forum for state-of-the-art developments and research in the theory and design of computers, as well as current innovative activities in the applications of computers. In contrast to other journals, this journal focuses on emerging computer technologies with emphasis on the applicability to real world problems. Current areas of particular interest include, but are not limited to: architecture, networks, intelligent systems, parallel and distributed computing, software and information engineering, and computer applications (e.g., engineering, medicine, business, education, etc.). All papers are subject to peer review before selection.

A. Procedure for Submission of a Technical Paper for Consideration

1. Email your manuscript to the Editor-in-Chief, Dr. Fred Harris, Jr., Fred.Harris@cse.unr.edu.

2. Illustrations should be high quality (originals unnecessary).

3. Enclose a separate page (or include in the email message) the preferred author and address for correspondence. Also, please include email, telephone, and fax information should further contact be needed.

B. Manuscript Style:

1. The text should be **double-spaced** (12 point or larger), **single column** and **single-sided** on 8.5 X 11 inch pages.
2. An informative abstract of 100-250 words should be provided.
3. At least 5 keywords following the abstract describing the paper topics.
4. References (alphabetized by first author) should appear at the end of the paper, as follows: author(s), first initials followed by last name, title in quotation marks, periodical, volume, inclusive page numbers, month and year.
5. Figures should be captioned and referenced.

C. Submission of Accepted Manuscripts

1. The final complete paper (with abstract, figures, tables, and keywords) satisfying Section B above in **MS Word format** should be submitted to the Editor-in-Chief.
2. The submission may be on a CD/DVD or as an email attachment(s) . **The following electronic files should be included:**
    - Paper text (required).
    - Bios (required for each author). Integrate at the end of the paper.
    - Author Photos (jpeg files are required by the printer, these also can be integrated into your paper).
    - Figures, Tables, Illustrations. These may be integrated into the paper text file or provided separately (jpeg, MS Word, PowerPoint, eps).

3. Specify on the CD/DVD label or in the email the word processor and version used, along with the title of the paper.

4. Authors are asked to sign an ISCA copyright form (http://www.isca-hq.org/j-copyright.htm), indicating that they are transferring the copyright to ISCA or declaring the work to be government-sponsored work in the public domain. Also, letters of permission for inclusion of non-original materials are required.

Publication Charges

After a manuscript has been accepted for publication, the contact author will be invoiced for publication charges of **$50.00 USD** per page (in the final IJCA two-column format) to cover part of the cost of publication. For ISCA members, $100 of publication charges will be waived if requested.

January 2014