
Andrzej Nowak


Paul Lukowicz


Paweł Horodecki

Assessing Artificial Intelligence for Humanity

Will AI be Our Biggest Ever Advance — or the Biggest Threat?


ISTOCK/LAGARTOFILM

Recent rapid advancements in Artificial Intelligence (AI) are arguably the most important dimension of humanity's progress to date. As members of the human race, that is, *homo sapiens*, we are defined by our capacity for cognition. Until now, humans were the only species capable of higher cognitive functions. But today AI has advanced to a stage where on many cognition-related tasks it can match and even surpass the performance of humans. Examples include not only AI's spectacular successes in winning Go [1], chess [2], and other board games with humans, and in surpassing humans on fully defined world puzzles. But AI is also now achieving extremely high efficiency in practical applications such as speech and object recognition, self-driving cars, intelligent tutoring systems, efficient decision support systems, and in the capacity to detect patterns in Big Data and in constructing accurate models of social behavior.

Thus, for the first time in history, we must ask ourselves: "has our monopoly on intelligence, however defined [3], been challenged?"

The development of AI is vigorously supported by industry. AI can radically cut costs for industry and corporations by reducing paid employment of humans. AI can also enhance the quality of products and services, allowing the scaling up of activities, and assisting in developing new products and services. Using AI, humans can potentially, on the one hand, solve some of our most challenging practical problems, and at the same time provide labor to replace humans at boring jobs. On the other hand, AI can reach levels of intelligence never before attainable for humanity.

Concurrent to AI's development and support from industry, there also recently has been much discussion on whether Artificial Intelligence could perhaps become the seed of the destruction of humanity [4], [5]. The potential danger of AI became one of the key topics of public discourse with the publication in January 2015 of a letter initiated by Stephen Hawking and Elon Musk, and signed by many prominent AI researchers [6]. In the discourse Musk called AI research "summoning the

Digital Object Identifier 10.1109/MTS.2018.2876105
Date of publication: 30 November 2018

demon," and Hawking warned that the development of AI could "spell the end of human race."

The apparent question is then, which of these visions is correct – if any? Should the development of AI be supported by all means as the biggest promise to humanity? Or should we try to stop it because it represents the ultimate danger?

Our answer is that, as often in life, it's more about shades of grey than about black and white. The question is not whether we should develop AI or not, but rather what sort of AI we should develop.

What all the "dark" scenarios have in common is a view that AI is being developed independently of human cognition and as an alternative to it. In these scenarios, AI develops its own highly efficient systems of knowledge and reasoning that are largely incompatible with the human way of thinking and acting. Thus AI cannot really support humans and collaborate with them. Instead, because of their efficiency, AI systems gradually replace humans in more and more tasks. As a consequence, we see increased loss of human control over the world and loss of human autonomy. In the "dark" scenarios AI actively turns against humans. Although these scenarios are quite diverse, the end point of them is similar: humans lose their leading role as intelligent beings, lose control of the world, lose their freedom and, in some of these scenarios, the physical existence of humanity is threatened.

In the various "grey" scenarios there is no direct action by AI to harm humans. Instead there is a gradual loss of autonomy and control resulting from more and more tasks and decisions being delegated to AI systems. A particularly dangerous aspect of such scenarios is "chaos like" emergent phenomena resulting from unforeseeable interactions between AI systems devoted to different tasks. Thus, while each AI may actually be able to perform its own tasks better than a human (at least from a short-term perspective), without proper coordination and without the ability to consider the big picture, the sum of the actions of all AI may lead to negative if not catastrophic long-term consequences.

The general concept is something that anyone who has used a car navigation to bypass a traffic jam knows. While each individual navigation system makes a recommendation that seems to make sense from an individual point of view, the sum of decisions made by many navigation systems may lead to the creation of new traffic jam on the alternative routes.

Another hotly debated point is question of whether AI can have the ability to replicate human creativity, intuition, and inventiveness, which are crucial in dealing with fundamentally new situations.

Positive AI Development Scenario

We propose that the development of AI can take a radically different route, which can be termed Human-Centered AI. Human-Centered AI focuses on collaborating with humans, enhancing human capabilities, and empowering humans to better achieve their goals. In other words, the well-being of humans is the superordinate goal of the development of AI.

In a positive AI development scenario, AI can be the biggest accomplishment in the cognitive development of the human race. AI sensing can dramatically extend the information acquisition capacity of humans. AI can store and integrate the knowledge of the human race. Machine-learning techniques can generate knowledge previously inaccessible to humans. AI can interconnect humans in new ways and optimize functioning of techno-human systems. AI can collaborate with humans and support them in the service of human goals. Because in this scenario AI supports humans in making more informed decisions rather than making decisions for humans, the results will combine the strengths of AI with human strengths such as creativity, innovation, and intuition (qualities that today we do not know how to replicate in AI systems).

Human-Centered AI needs to understand human lines of reasoning, and relate to human morals, motivations, and emotions in that reasoning.

Thus, in the positive scenario, the results will go beyond the direct extrapolation of previous experience and allow fundamental changes. It will therefore be possible to develop solutions to radically new situations. This scenario will happen if AI is used to enhance human intelligence, rather than to replace it. In other words, the positive scenario will happen if AI is used to acquire, generate, integrate, access, and process the knowledge of human race, rather than to develop an alternative system of knowledge and its representation that is inaccessible to humans.

In this article we reflect on possible scenarios of AI development, and how they are related to the type of AI that is being developed. That is, is the AI being built a "Function-Oriented AI" or a "Human-Centered AI." We consider the differences between these two approaches to AI development and outcomes. We

also argue that although, in the long run, Human-Centered AI is likely to be superior to Function-Oriented AI, Human-Centered AI needs more support in its early stages of development in order to have a fair chance of prevailing in the competition with Function-Oriented AI.

The real danger of AI lies not in sudden apocalypse, but in the gradual degradation and disappearance of what make human experience and existence meaningful.

Creation of a Vicious Super Mind

The most catastrophic "dark" scenario envisioning AI apocalypse is the creation of a super-intelligence surpassing that of humans that will rapidly advance by accessing Big Data, super strong learning algorithms and positive feedback loops created by self-improving AI architecture. As soon as AI significantly surpasses human intelligence it becomes a natural competition to humanity. The natural features of advanced AI systems according to this scenario, likely lead to the realization of this dark outcome. The principle of self-preservation would lead to the development of defensive strategies on the part of the AI. These AI defensive strategies, such as hiding itself, replication, and resource maximization would result in competitive behaviors, and rapid physical expansion (4).

If AI can take control of military applications, for example automated weapon systems, it would be in position to wipe out humanity in its bid for self-preservation and expansion. In a scenario depicted by the movie "Her" and discussed in popular books (7), the creation of a super intelligence would result in the emergence of a super powerful mind equipped with awareness and its own goals that are likely contradictory to the goals of humanity. While this spectacular scenario has captivated public attention, from a technological point of view there is no plausible path from today's state of the art to the vicious super-mind. Today's AIs are all Turing machines and learning is mostly about complex statistical modeling and optimization on huge data sets, while reasoning is about information representation, efficient search, and clever application of complex rule systems. Clearly, we cannot prove beyond doubt that sufficiently complex statistical analysis of sufficiently large data sets will not lead to the emergence of consciousness. But there is no solid scientific evidence indicating that such an emergence of consciousness could happen, or how it could do so.

Emergence of a Global Socio-Technological Quasi-Mind

While not completely ruling-out the likelihood of this futuristic scenario of the creation of a vicious super mind, we argue that the danger of AI may come in a somewhat different, more "grey" form. We also argue that this scenario is likely to be already occurring. In this "grey" scenario the danger is not an apocalypse of physical elimination of the human race by an alternative, superior, artificial, self-aware mind. Rather the danger lies in the gradual disappearance of what makes us human, and of what makes our existence meaningful. We are talking here about our human ability to make choices for the realization of our needs and values, and our capacity to be subjects of our existence in pursuit of self-realization. Instead, humans may gradually became passive elements of an emerging global socio-technical system — a system composed of machines, algorithms, sensors and actuators, AI programs, and humans interacting in the globally present Internet, and Internet that is ever-present due to mobile technologies and ambient intelligence.

In this grey perspective, the critical question concerns the degree to which individuals, who are elements of this global system, have freedom of choice and action, and to what degree are they "enslaved" by the system. Are the internal processes of a human, in effect, dictated by the arising, global social-technical system, its algorithms and its emergent processes and goals (or more accurately, quasi-goals, understood as standards of regulation)?

In this negative scenario, humans are losing their freedom and becoming elements that process information in the service of the global techno-social system. The essence of this question is: what are the real chances for humans to break-out of the choices dictated by the system of which they are an element? What are their chances to retain the capacity for independent, critical thinking? Can they retain emotions and feelings dictated by their internal processes, rather than those dictated by information tailored to manipulate their emotions, or by autonomous decision-making?

To what degree do these processes serve humans' true needs, values, and goals versus those of the techno-social global super-computer? Another dimension of this question is "to what degree do interactions and contacts between individuals retain a human character, characterized

by the intrinsic value of the inner experiences of other humans?" The practical question is "how can humans retain their autonomy and free will amid the emergence of the global techno-social system?"

Some elements limiting human choices as a result of interaction with the global techno-social system are already visible. For example, bias introduced in search algorithms to match information collected from past searches on interests and views, reinforces existing views and patterns of individual decisions. This limits the capacity for innovative choice, leading to increasingly polarized opinions and behavior (8). Moreover, recommendation algorithms based on deep learning that follow viewers' interests, likely lead to distorted view of reality, where content that is divisive, sensational, and conspiratorial may promote fake news over objective journalism (9).

These choices and behaviors of individuals are increasingly controlled by sophisticated social influence mechanisms like micro-targeting (10). The limitation of autonomy may be the intended result of marketing efforts or political campaigns. It may also, however, be an emergent property of various algorithms interacting with each other and with other humans. Regardless of the source, the techno-social system may go in directions that do not serve an individual's goals, or those of the wider society.

Scenarios Leading to the Emergence of an Uncontrolled Techno-Social Quasi Mind

The question is, what are the scenarios leading to the emergence of a new AI-like meta-level system created by the interaction of nature, society, and technology? What is the likelihood of the occurrence of these scenarios, and will we recognize the rise of such a meta-system? In the most likely scenario, the super system will be created by the interaction of human cognition and computer information processing, which combines sophisticated information processing and extremely efficient AI learning algorithms on one side, with randomness inherent in human information processing on the other. The emerging information processing structure will likely be difficult, or even impossible to design or even to be understood by humans. Inconsistency and contradiction inherent in human decisions and actions is likely to amplify the complexity of the emerging system.

Such a socio-technological system would not need to follow strict rules of rationality or to strictly obey well-known economic rules of self-improving systems (4). In particular, it might exhibit deviations from the optimization of the consumption of resources, which could potentially be catastrophically dangerous to nature and the natural environment. On the other hand, having as its elements AI components intentionally designed to be capable of formulating distant goals, such a system is likely to have emergent functions or distant goals that may not even be known to society, at least not until severe consequences are visible.

With the rise and self-organization of such a quasi-mind system, the AI-like system is likely to increase its share of the control of information processing and decision making at the expense of humans. One reason is that humans have limited resources for processing capacity, attention, memory, etc. Moreover, humans get tired of making decisions. This has been described as *decision fatigue* (11) in economics and psychology (12), but also as *ego depletion* (13). Individuals are thus expected to have a drastically decreased capacity to be the source of independent influence on a socio-technical network. With growth of the size of the network, the number of interactions required to maintain any significant level of control of the network can easily exceed human capacity for interactions (14) and choice.

We believe that this can be formally proven or demonstrated using Bayesian (causal) networks (cf., (15)), or complex systems and chaos theory, or quantum analogs. In such networks, beyond some threshold number of connections, the collective phenomena in the network would emerge, and the network behavior would become, in a sense, uncontrollable to the nodes. What is even more important is that it is possible that, without being aware of it, humans may just function as procedures or subroutines of the larger system, being part of the higher-level computational process of the socio-computational system. In this role humans may generate new goals without even being aware of it, analogous to "games designed on purpose" (16), where a byproduct of playing the computer game is supposed to be a solution of some problem. The major difference is that the user would have only a little (if any) knowledge of what the game is that he or she is taking part in.

It is also not clear how human intelligence might evolve in this scenario. The impacts of such a techno-social system might be that people would lose their independence, and also experience a deterioration of intelligence. Or it might be that at least some aspects of human intelligence might increase, for example as an element of subroutines that would suit the AI-system's distant goals. It is more likely that, in general, humans' long-term memory function might decrease (e.g., an analysis of Google books suggests that, as a society and culture, we are forgetting the past faster (17)).

Let us stress that, even if a standard design AI were apparently relatively well controlled by humans (in

the sense that it would not attack humanity directly, via spectacular s-f type extinction, etc.), the emergence of some AI-type human-network system might still occur as a by-product (e.g., via the Internet of Things) that would benefit from human intelligence and partial randomness and/or free will as a resource. This might formulate the distant quasi-goals that would influence the Nature-Society-Technology triangle.

The final quasi-goals or output states of the process may be completely unpredictable due to 1) nonlinearity of the process, 2) computer power becoming different in some way from what we currently understand computing power to be, or 3) the human component that is involved in a nonstandard way. For example, what if humans that were originally designers of the system's algorithms then have their behaviors "designed" in a sophisticated back-reaction. The possible loss of human independence might occur gradually, or through a specific transition, beyond which there is a point of no return. This might involve an energetic breakdown, irreversible genotype changes, evolutionary changes in individual or social behavior that would lead to paradoxical losses, some forms of addiction, or something else.

The fundamental question is how to recognize that the emergence of such system is happening. By its very definition, if the capacity of the system goes beyond that of human beings, then the distant goals should be unpredictable, and only visible *a posteriori* to humans. However, there might be some signal-type behaviors. An era of computational activity of an AI system might, in addition to convergence to distant goals, lead to some new by-product regularities or repeatable phenomena in its functioning.

What would differentiate activities of the artificial systems from other natural regularities coming from direct human activities (e.g., new highways)? First, AI-enhanced mobile communications, social networks, and digital media increase the density and strength of connectivity between humans. Such densely connected systems tend to be less stable and more prone to cascading effects [18]. Second, the speed with which things happen when connected AI systems are involved makes it difficult for humans to react to problems. Finally black-box-like machine learning systems (e.g., the AlphaGo Zero self-thought program) produce solutions whose bases are mysterious — i.e, where it is very difficult to discern how it was possible to find the solution.

Another signature of the development of the far distant goal or meta–structure might be visible at the level of resources — especially at higher energy consumption (since it is always necessary to keep complex systems far from equilibrium, and a possible distant quasi-goal might have extra complexity). This however applies to the complex emergent system, but does not need to apply to the computation process (see the concept of reversible computation [19]). However, there is a chance then that the reversible character may be tracked without referring to the energy resource (with time as a natural resource). It may be that the AI-type socio-technological system activity would gradually lead to the loss of free will, by narrowing the perspective so that society would see some processes (behaviors) as unavoidable, before they actually were so.

Decline of Cognitive Capacities

The rapid development of AI that replaces rather than augments human intelligence can also dramatically diminish the capacity for cognition of the human race. In this scenario, human information processing is delegated to AI, and humans just get answers. They don't gain understanding of the knowledge and processing rules that led to the solutions. Deep learning algorithms [20] provide an example of AI systems, that if provided with enough learning examples, processing power, and time can learn almost any pattern, classify objects, predict next events, and make decisions that are on average better than those made by humans. When the tasks are routine, statistics show the superiority of artificial neural networks over human performance. Lower costs and statistically better performance of neural networks over human experts raise the temptation to replace human judgment and decision-making with neural networks not only for simple tasks, but also for complex decision making and judgment tasks such as employment decisions and political and business strategic choices.

Although replacing human cognition by AI may, in some instances, have spectacular short-term advantages, it can be disastrous in the long term. Today's machine learning systems create abstract representations that are alien and mostly inaccessible to humans. This type of abstract knowledge cannot be blended with the existing knowledge of humans. As a consequence, while artificial neural networks can replace humans in routine tasks, they do not produce knowledge that can be used by humans, and they do not add to the knowledge possessed by the human race. On the contrary, AI that replaces humans presents a grave danger to the cognitive skills of the human race. It is a most threatening factor that could cause rapid decay and decline of human cognitive skills.

Any skill that is not used decays. As an increasing range of tasks is delegated to AI, humans will lose the knowledge and skills to perform these tasks and will become helpless without AI. This, in an increasing positive feedback loop, will cause an increasing tendency to delegate all the difficult tasks to AI. Moreover, since humans cannot understand the bases for the decisions made by AI, they will increasingly lose control of the processing of information. Trusting AI will become the only choice — without the possibility of checking if the AI decisions are beneficial for humans. This can become disastrous in several ways. Most importantly, it can lead to the decline of human competencies and cognitive skills. Skills that are not used can diminish to a catastrophic extent. By delegating information processing to systems that use rules that humans do not understand, humanity will lose a significant degree of the cognitive competencies that have given us the adjective "sapiens." Delegating information processing also implies that when novel creative solutions are required due to changed conditions or new opportunities or threats, humanity may be helpless because algorithms trained on existing data cannot cope with radically novel situations. So, while most of the time AI may outcompete humans, in the most critical situations it will fail with possibly disastrous consequences.

In summary, our species, *homo sapiens*, is defined by our capacity for cognition. Rapid development of AI can change this capacity in a most profound way, for the better or worse. In one scenario, by enhancing human cognitive capacity AI can elevate humanity to an unprecedented, or even unforeseen, level, of perception, knowledge, understanding, and reasoning. In another, it can take away what makes us human by effectively diminishing our cognitive capacity to acquire useful information, and to diminish our knowledge and our capacity to reason. It can wipe out our understanding of the world around us.

The direction in which the development of AI will take us depends on what kind of AI we develop. The rapid development of AI can either make us more human, or alternatively, can become the biggest existential threat to humanity [21]. This threat is not only in the sense of physically eliminating us from the face of the Earth, as some scenarios predict, but in a much less spectacular way by reducing our cognitive capacity and, in effect, taking away our humanity, the adjective *sapiens* that defines our species. While delegating human reasoning and decision making to AI may trigger different negative scenarios, Human-Centered AI is likely to result is the transition to a higher level in the development of intelligence of human race. The critical question then may be how to develop AI, and the global super-net, so as to leave space for free will, free choice, and the self-realization of individual humans.

Human-Centered AI

The common element in all the negative scenarios is that AI develops its own knowledge system that is inaccessible to humans. AI develops decision rules that are oblique to human understanding, and AI is focused on replacing rather than supporting humans.

In contrast, Human-Centered AI aims to interface and extend human capabilities [21], enhance human decision making, and serve human goals on both the individual and societal level. Human-Centered AI is designed along ethical and value-oriented principles that are not an optional "add-on" or a "by design" feature. The concept of Human-Centric AI envisions future AI technology that will synergistically work with humans for the benefit of humans and human society:

Artificial Intelligence that replaces rather than augments human intelligence can dramatically diminish the capacity for cognition.

- Instead of replacing humans we need to focus on enhancing human capabilities allowing people to improve their own performance and successfully handle more complex tasks.
- Instead of prescriptive systems telling people what to do we need to focus on systems that empower humans to make more informed decisions and help them harness and channel their creativity.
- Instead of creating unpredictable "black box" systems we need to focus on explainable, transparent, validated, and thus trustworthy systems optimally supporting both individuals and society as a whole in dealing with the increasing complexity of a networked, globalized world.
- We need to include values, ethics, and privacy as core design considerations in all of our AI systems and applications.

For the above vision to become reality a large scale. long term research effort is needed that goes from the underlying fundamental unsolved problems of AI, through specific novel technologies in different applied AI domains to making broad impact in

relevant socio-economic areas. Such an effort must bring together three main communities: research, industry, and societal stakeholders. We are currently pursuing this vision in the Human^{E} AI initiative (www.human-ai.eu).

What does this mean in concrete terms? Consider a judge, doctor, policy maker or manager facing a complex decision that has to be made on the basis of a large, noisy data basis and involves a variety of aspects that may not all be within the core competence of the decision-maker. Since such decisions often have grave personal and/or social consequences and include complex ethical and emotional aspects, a complete replacement of human decision makers by AI is clearly undesirable, even if it were feasible. Existing decision support systems are mostly about guiding a person through a predefined decision tree, which means that while the decision may formally be taken by the human, it is often largely pre-determined by the system. Data mining and analytics systems leave much more freedom to the user, at the price of a potential information overload.

Human Centric AI should be able to *truly debate problems* with the human user. A Human-Centered AI needs to understand human lines of reasoning, and relate to human motivations and emotions and to the moral assumptions and implications in that reasoning. The AI needs to help the human partners, challenge their assumptions, and to provide and explain alternative ways of seeing a problem (given the AI's particular analytical abilities and data access). Only an AI system that is capable of such a rich and reflective discussion with a human can optimally support informed decision-making while leaving sufficient space for human intuition, inventiveness, and creativity.

Implementing such systems is related to two well known fundamental "Grand Challenges" of AI. First is the ability to build and maintain comprehensive, differentiated world models. A key aspect of human intelligence is a world model based on a huge amount of experience. The human world model is based on complex, often ambiguous semantics, and on a dense web of associations that allow a variety of levels of implicit communication (including irony and figurative speech). These factors are the basis for human creativity and inventiveness. Although achievements such as the IBM Watson's, or the Debater project success at Jeopardy, are great advances towards more advanced AI world models, we are still very far from the comprehensiveness, richness, and subtlety of the human world understanding.

The second related challenge is the explainability of machine learning models. Thus, many recent AI success stories are based on the application of complex statistical analysis to massive amounts of training data. As powerful as such methods have proven to be, they have the disadvantage of being very hard (often impossible) for humans to understand and interpret, making AI-based decisions difficult for humans to accept [22]. This goes against the vision of AI systems that can debate with humans and synergistically work together with people, including learning from experts.

As an example of the differences between the two approaches, in the Function-Oriented AI approach, a company may use a deep learning algorithm to make personnel decisions. A deep learning network, based on the past history of productivity of workers with different characteristics, would develop its own algorithm that would be encoded in connection to strength between nodes of the network. This algorithm would not be accessible to humans.

If the company adopts this algorithm, it would make personnel decisions in a way that no one in the company understands. Moreover, because this algorithm would reflect only past experiences, it would be likely to fail if the business environment changes.

In contrast, Human-Centered AI would analyze a huge amount of data about worker productivity, and would reveal complex patterns underlying that productivity to managers. This knowledge could be used to formulate rules that would underlie hiring and firing decisions in the company. These rules could be revealed to workers, and if needed implemented into software for automated decision making. The decision-making software could be changed by managers proactively in anticipation of planned changes in the company's strategy.

As another example, an AI-based recommendation system, based on a deep-learning algorithm that has been trained to maximize the time users spend on a social media site, will recommend to users content using rules that may be not understood by anyone. As evidenced by prior experience (e.g., [3]), this algorithm is likely to develop rules promoting highly distorted content. The distorted content is then disruptive to constructive social processes, which violates the values of society. A high number of human moderators will be needed to neutralize the bias introduced by the algorithm.

In contrast, Human-Centered AI would use sophisticated algorithms to discover which content is most attractive. These findings would add to already existing knowledge. While constructing algorithms for the recommendation systems, Human-Centered AI would also take into account societally accepted values

such as trustworthiness of the information, and avoidance of hate.

Dissatisfaction with AI that operates on the basis of knowledge that cannot be understood by humans has resulted in a new strong movement towards the concept of "explainable AI" — or XAI. The XAI concept has gained popularity in science [23], business [24], [25], and military circles [26]–[27] working with AI. The goal of this trend is to decipher the knowledge and decision algorithms used by machine learning applications and translate this into rules and knowledge accessible to humans. Understanding the rules of AI increases trust in, and accountability for, the control and safety of AI applications based on machine learning [24]. Because most machine learning algorithms are inherently complex, however, full explanation of the algorithm may be impossible. The goal of the approach may then be to offer a very simplified explanation to humans, (for example, to name the most important factor in the decision), rather than to extract the maximal knowledge of the algorithm used by AI because people prefer simper explanations [28]. Although the approach of XAI gives humans some control, it usually assumes that AI systems can make better decisions than humans; thus it tends to delegate decisions to AI.

Another key aspect of Human-Centered AI is the ability to act and interact in complex social contexts [29]. As an example, consider a well-known, seemingly trivial problem: automatically deciding if, given a certain setting and a specific caller, if it is appropriate for a mobile phone to ring or not. At the core of the problem is the need for an in-depth understanding of the fine points of the social context in which the user is currently situated and the ability to anticipate the potential significance of taking or delaying the call in the framework of the user's life. Furthermore, the user's current activity, mood, and priorities must be taken into account. In the same meeting it may or may not be appropriate to ring depending on who is currently present, who is talking, and how the meeting has evolved. Thus, if the user is about to successfully convince his bosses of something he deeply cares about, then taking a call is not helpful. On the other hand, if he is going to lose the argument anyway, then the call may be more important. A call from the same person may have a very different priority depending on recent interactions and current expectations or intentions towards the caller. While research areas such as context-aware computing [30], affective computing [30], or social computing [31] have considered various aspects of the problem, acting and interacting within complex social settings and taking into account the full complexity of human feelings and decision making processes is another unsolved AI Grand Challenge.

Human-Centered AI could also enhance the functioning of human groups and socio-technical systems. This could be achieved by facilitating interactions between humans and technology through facilitating more productive social interactions, helping to find more trustworthy sources of information, helping to delegate information processing and decision making to most qualified individuals, and by providing on-the-go knowledge facilitating group productivity [33]. In contrast to Function-Oriented AI, the rules governing these processes would be accessible to and modifiable by humans [32].

A final consideration in the design of Human-Centered AI systems is the integration of ethical values and social norms [33]. As AI systems influence more and more areas of our lives, their actions must be aligned well with the values and expectations of both users, and society in general, to be acceptable and accepted. This is a problem that goes beyond a mere technical integration and representation of ethical concerns and social norms within an AI system. It involves enabling the system to perform often inherently ambiguous ethical reasoning (which by itself is an open research problem). In addition, a well-informed discussion is needed among all stakeholders — researchers, industry, and the wider society — about the relevant ethical values and norms that AI systems should follow and under what conditions. Whatever values end up embedded into AI systems, it is essential that the design decisions about what is included be explicit and visible. People should be able to inquire and understand the underlying values that an AI system is optimized for.

Human^{e} AI Project

In summary, the concept of Human Centric AI envisions future AI technology that will synergistically work with humans for the benefit of humans and human society, focusing on enhancing and empowering humans rather replacing and controlling them. Core concerns are accountability, explainability, appropriate interaction concepts, and the inclusion of values, ethics, and privacy as core design considerations.

For the above vision to become reality, a large-scale long-term research effort is needed that goes from the underlying fundamental unsolved problems of AI, through specific novel technologies in different applied AI domains, to making broad impact in relevant socio-economic areas. Such an effort must bring together three main communities: research, industry, and societal stakeholders. We are currently pursuing this vision in the Human^{E} AI initiative (www.humane-ai.eu) and the European CLAIRE (https://claire-ai.org/) network of AI laboratories.

How the development of AI affects the cognitive capacities of humanity will depend on which route humanity adopts for the development of AI. If AI develops in a way in which its functioning in higher cognitive tasks is based on knowledge inaccessible to humans, and the main goal of AI is to replace humans in tasks requiring complex cognition, then the consequences may be disastrous. If, however, AI takes the Human-Centered approach, if AI champions the concept that its main goal is to extend human cognitive capacities and to generate knowledge accessible to humans (serving the goals as defined by Human-Centered AI) — then the development of AI may be the most significant achievement in the evolution of humanity.

Acknowledgment

This work was supported by funds from Polish National Science Center (project no. DEC- 2011/02/A/HS6/00231).

Author Information

Andrzej Nowak is the Director of the Center for Complex Systems in The Robert B. Zajonc Institute for Social Studies, and Professor at the Department of Psychology University of Warsaw, Warsaw, Poland; email: andrzejn232@gmail.com.

Paul Lukowicz is Full Professor of AI at the Technical University of Kaiserslautern in Germany where he heads the Embedded Intelligence group at Deutsches Forschungszentrum für Künstliche Intelligenz, (DFKI); email: Paul.Lukowicz@dfki.de.

Paweł Horodecki is professor at University of Gdańsk, International Centre for Theory of Quantum Information Technologies and Gdańsk University of Technology, Faculty of Applied Physics and Mathematics, also associated with the National Quantum Information Center of Gdańsk, Poland. Email: pawel.horodecki@pg.edu.pl.

References

[1] C. Koch, "How the computer beat the Go master," *Scientific American*, vol. 19, 2016.

[2] F.-H. Hsu, *Behind Deep Blue: Building the Computer that Defeated the World Chess Champion*. Princeton, NJ: Princeton Univ. Press, 2004.

[3] A.M. Turing, "Mind," *Mind*, vol. 59, no. 236, pp. 433–460, 1950.

[4] S.M. Omohundro, "The nature of self-improving artificial intelligence," presented at *Singularity Summit*, 2007.

[5] S. Hawking, S. Russell, M. Tegmark, and F. Wilczek, "Transcendence looks at the implications of artificial intelligence – But are we taking AI seriously enough?," *Independent*, May 1, 2014; https://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-9313474.html.

[6] Future of Life Institute, *Research Priorities for Robust and Beneficial Artificial Intelligence;* https://futureoflife.org/ai-open-letter, accessed Oct. 8, 2018.

[7] J. Barrat, *Our Final Invention: Artificial Intelligence and the End of the Human Era*. Macmillan, 2013.

[8] R. Epstein and R.E. Robertson, "The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections," *Proc. National Academy of Sciences*, vol. 112, no. 33, pp. E4512–E4521, 2015.

[9] P. Levis, "How YouTube's algorithm distorts truth," *The Guardian*, 2018; https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth.

[10] G. Farnadi, G. Sitaraman, S. Sushmita, F. Celli, M. Kosinski, *et al.*, "Computational personality recognition in social media," *User Modeling and User-Adapted Interaction*, vol. 26, no. 2–3, pp. 109–142. 2016.

[11] M.A. Boksem and M. Tops, "Mental fatigue: Costs and benefits," *Brain Res. Rev.*, vol. 59, no. 1, pp. 125–139, 2008.

[12] E. Polman and K.D. Vohs, "Decision fatigue, choosing for others, and self-construal," *Social Psychological and Personality Science*, vol. 7, no. 5, pp. 471–478, 2016.

[13] R.F. Baumeister, E. Bratslavsky, M. Muraven, and D.M. Tice, "Ego depletion: Is the active self a limited resource?," *J. Personality and Social Psychology*, vol. 74, no. 5, p. 1252, 1998.

[14] R.I.M., Dunbar, "Neocortex size as a constraint on group size in primates," *J. Human Evo.*, vol. 22, p. 469, 1992.

[15] R.M. Shiffrin, "Drawing causal inference from Big Data," *PNAS*, vol. 113, no. 27, pp. 7308–7309, 2016.

[16] L. von Ahn and L. Dabbish, "Designing games with a purpose," *Commun. ACM*, 2008; http://www.cs.cmu.edu/~biglou/GWAP_CACM.pdf.

[17] J.B. Michel *et al.*, "Quantitative analysis of culture using millions of digitized books," *Science*, 2011.

[18] D. Helbing, "Globally networked risks and how to respond," *Nature*, vol. 497, no. 7447, p. 51, 2013.

[19] C.H. Bennett, "The thermodynamics of computation — A review," *Int. J. Theoretical Physics*, vol. 21, no. 12, pp. 905–940, 1982.

[20] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning* (Adaptive Computation and Machine Learning series). Cambridge, MA: M.I.T. Press, 2016.

[21] https://www.theguardian.com/technology/2014/oct/27/elon-musk-artificial-intelligence-ai-biggest-existential-threat

[22] P., Lukowicz, S Pentland, and A. Ferscha. "From context awareness to socially aware computing." *IEEE pervasive computing* 11.1 32–41, 2012.

[23] Lepri, B., Oliver, N., Letouzé, E. *et al.* Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philos. Technol.* (2017). https://doi.org/10.1007/s13347-017-0279-x

[24] T. Miller, P. Howe, and L. Sonenberg, "Explainable AI: Beware of inmates running the asylum," in Proc. IJCAI-17 Workshop on Explainable AI (XAI), Aug. 20, 2017; http://www.intelligentrobots.org/files/IJCAI2017/IJCAI-17_XAI_WS_Proceedings.pdf#page=36.

[25] "Explainable AI," *pwc UK*, 2018; https://www.pwc.co.uk/services/audit-assurance/risk-assurance/services/technology-risk/technology-risk-insights/explainable-ai.html.

[26] "The challenges and opportunities of explainable AI, *intel AI*, Jan. 12, 2018; https://ai.intel.com/the-challenges-and-opportunities-of-explainable-ai/.

[27] D. Gunning, "Explainable Artificial Intelligence (XAI)," *DARPA*, https://www.darpa.mil/program/explainable-artificial-intelligence, accessed Nov. 2018.

[28] T Lombrozo, "Simplicity and probability in causal explanation," *Cognitive Psychology*, vol. 55, no. 3, pp. 232–257, 2007.

[29] B. Schilit, N. Adams, and R. Want, "Context-aware computing applications," in *Proc. 1994 Workshop Mobile Computing Systems and Applications*, pp. 85–90. IEEE, 1994.

[30] C.L. Lisetti, "Affective computing," *J. Pattern Analysis and Applications*, vol. 1, no. 1., Mar. 1998.

[31] F.Y. Wang, K.M. Carley, D. Zeng, and W. Mao. "Social computing: From social informatics to social intelligence," *IEEE Intelligent Syst.*, vol. 22, no. 2, 2007.

[32] A. Schmidt, "Augmenting human intellect and amplifying perception and cognition," *IEEE Pervasive Computing*, vol. 16, no. 1, pp. 6–10, 2017.

[33] A. Nowak, R. Vallacher, A. Rychwalska, and M. Kacprzyk, "The target in control," submitted for publication, 2018.

[34] J. Van Den Hoven and J. Weckert, Eds., *Information Technology and Moral Philosophy*. Cambridge, U.K.: Cambridge Univ. Press, 2008.