# Artificial Intelligence as a Pathway to Our Future

*Jiří Wiedermann*

*Jan van Leeuwen*

Technical Report UU-PCS-2024-01
January 2024

Philosophy of Computer Science
Department of Information and Computing Sciences
Utrecht University, Utrecht
The Netherlands

Series: UU-PCS

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

# Artificial Intelligence as a Pathway to Our Future[1]

Jiří Wiedermann[2] and Jan van Leeuwen[3]

**Summary** Generative AI is the talk-of-the-town. Should it be welcomed or should it be feared? Is it a great extension of the human mind that should be cherished or should its development and use be limited? This essay aims to shed light on these tantalizing questions from a general, philosophical perspective. Why is AI being developed, and what is the purpose of using it? What does it mean for mankind, for our future? The answers can be traced in the development of the field to date and explain both the urge to develop it further and the urge to control it.

**Keywords**: AI regulation, artificial wisdom, DIKW hierarchy, epistemic computation, generative AI, illusory intelligence, knowledge, large language models, 4E cognition.

## 1 Introduction

Consider this *riddle*: what is interactive and wise, yet not alive? Just such entities have recently appeared among us. No one knows what they look like, whether they have a body, senses, or how they reason. All we know is that they can ostensibly be reached in *cyberspace*. But it is quite clear to all who have interacted with them, or at least heard of them, that these entities can converse in many languages, are quite knowledgeable about virtually any topic, understand the vast majority of dialogues very well, and argue meaningfully. Unfortunately, occasionally they make things up that aren't necessarily true and it isn't always easy to tell when.

Of course, the entities we're talking about are instances of *generative artificial intelligence*. Its properties are so interesting and contradictory that the debate about it has attracted the attention of both the lay and professional public and is covered in all the news media. AI experts, neuroscientists, philosophers, cognitive scientists, linguists, prominent business figures, and intellectuals argue about whether generative AI is useful or dangerous to mankind, whether it can possess mental properties comparable to humans, and whether to develop it further at all, or under what conditions. Shall we embrace it, cultivate it, develop it, cooperate with it or, on the contrary, shun it, fear it, forbid it, and make no attempt to meliorate it?

When questioning whether the emerging possibilities of artificial intelligence are potentially harmful or dangerous to us, it makes sense to address the following deeper questions first: "*Why are we developing artificial intelligence?*" and "*What is the purpose of using artificial intelligence?*" These philosophical questions are now more important than ever before, as they may impact on the development of our society in even the nearby future .

The answers to our questions are straightforward if we ask about the development and purpose of AI systems that address some specific and known problems. However, if the question is about artificial intelligence that is general, in the sense that it

[1] Revised and updated translation of: J. Wiedermann and J. van Leeuwen, *"Umělá inteligence jako zdvižˇ do naší budoucnosti"*, in: V. Mařík, M. Trčka, and D. Černý (eds.), *"Proč se nebát umělé inteligence"*, Academia, Prague, 2024. The article is based on [12].

[2] Institute of Computer Science of Czech Academy of Sciences and Karel Čapek Center for Values in Science and Technology, Prague, Czech Republic, email: jiri.wiedermann@cs.cas.cz.

[3] Dept. of Information and Computing Sciences, Utrecht University, Utrecht, the Netherlands, email: J.vanLeeuwen1@uu.nl.

can solve any problem and make decisions on its own, the answer is more complex. In our case here, we are interested in answers to our questions that are based on a deeper understanding of the concept of artificial intelligence and its development to date. The answers should emerge from some concept that applies to "any AI", provide new insights into the nature of AI, and allow for an extrapolation of trends in AI, thus allowing for rational ideas about the future of AI. By "any AI" we mean both what we currently think of as artificial intelligence and all kinds of artificial intelligence in the future, on Earth and anywhere in the Universe.

Ideally, we want the answers to our questions to enable us to determine where artificial intelligence is going. But how can we tell where the field is headed? Can we identify a desirable course of further development as supported by the evidence? What consequences will it have for our future? Are there any obstacles or pitfalls we should avoid? Convincing answers to such "sub-questions" can only be given when underpinned by convincing facts.

Our line of reasoning will therefore be based on an appraisal of the historical and current trends in information technology, specifically in artificial intelligence, as well as on philosophical considerations and, in part, on formal theory. This will help us to trace the objective nature of the field and, within this framework, to propose not only well-founded answers but also an extrapolation of the importance of the development of artificial intelligence, and of generative AI in particular, for our future.

We will argue that future artificial intelligence will be qualitatively different from current AI, by using knowledge and experience to *act purposefully*. This will endow AI systems with a new advanced ability that can be summarized by the notion of *(artificial) wisdom*. In the context of AI systems, artificial wisdom is a relatively new concept that enhances the power of these systems so they can cope with the complexity, variability, and ambiguities of the real world that traditional systems could not. The respective systems will be able to learn and adapt over time and to solve complex problems. Last but not least, they will be able to make own, ethical decisions.

From this, we will further derive a vision of future AI obeying artificial wisdom that is appealing, plausible, and safe. We will explain why and how such AI can be seen as a pathway to our future, and under what circumstances and obligations one will not need to fear it.

## 2 Trends in the use of computing technologies

To identify trends in artificial intelligence, we draw on the trends in the development of computing technologies from the middle of the 20th century till the present. It is clearly depicted in Figure 1. We see how the computational power of computers has continuously grown in accordance with *Moore's law* and how the corresponding computational resources and paradigms have changed accordingly.

More interesting from the perspective of this essay, however, is the development of the type of data that is processed, and how it is processed, as a function of the increase in computing power over time. This is depicted in Figure 2. The two figures clearly show that, as information technologies and their application are steadily evolving, their "*information power*" is steadily advancing upward in the successive levels of the so-called DIKW hierarchy: *data, information, knowledge, wisdom*. The hier-

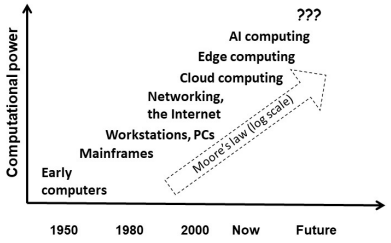


**Fig. 1.** Evolution of computing technologies

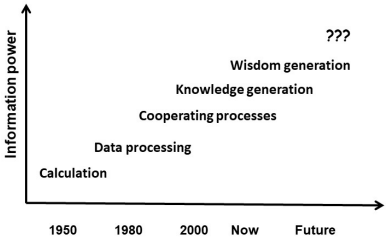


**Fig. 2.** Growth of information processing power

archy captures the fact that, typically, information is defined by data, knowledge by information, and wisdom by knowledge (and understanding).

The concept of the DIKW hierarchy was popularized in 1988 by Russell L. Ackoff in his Presidential Address to the ISGSR, now the *International Society for the Systems Sciences*. However, the idea had been around in various forms before among computer, control theory, and operations research professionals and can therefore be considered "folk wisdom." The DIKW hierarchy is often depicted as a pyramid, with data at the bottom level and wisdom at the top. The shape of the pyramid captures the fact that wisdom is the highest form of knowledge and that it can only be attained through data which transforms progressively from information and knowledge to wisdom. We assume that wisdom includes its own meta-level ("meta-wisdom").

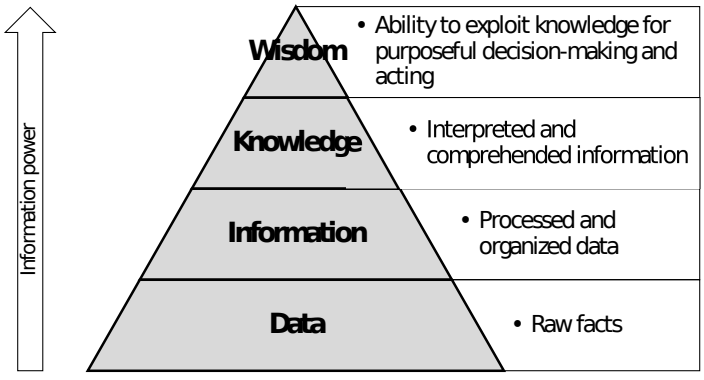


**Fig. 3.** Data transformation in AI systems

The DIKW hierarchy is actually nothing more than a rough sketch of a very general *architecture* of an information processing system. The individual boxes adjacent to the pyramid indicate what is being processed by means of the data, in a bottom-up direction, at that level. The transformation of data between levels of the pyramid is done using algorithms, often very complex ones (e.g. neural networks, statistical computation, pattern recognition, trend identification, etcetera), which depend on the desired outcome of the transformation and the type of data entering and exiting the

transformation process. More precisely, the data entering the pyramid can be numbers, texts, codes, images, outputs of various sensors, or other unprocessed raw facts. Information-level processing may involve statistical computations, pattern recognition, trend recognition, etc. At an even higher level of knowledge, it involves building concepts, discovering patterns, and finding relationships between different groups of information. Finally, at the level of wisdom, it is looking for ways to effectively apply knowledge to new situations. This is where, for example, neural networks that mimic our ideas about how the brain works may come in.

The qualitative difference between the data at two adjacent levels in the DIKW hierarchy represents the degree of understanding of the data at the lower level, expressed by higher-level means, always relative to the mission of the system in question. Clearly, such a degree measure of understanding increases in a bottom-up direction.

The scheme above is very general and applies to any artificial or natural intelligence system that generates wisdom, regardless of the environment in which it operates and its mission. It applies, for example, to an automatic door, a self-driving car, an autonomous missile defense control system, to generative AI systems, the brain, and even to a so-called "super-intelligence." At this level of abstraction, the scheme of operation of such systems is the same.

According to Ackoff, the first three levels in the set-up can be "logically specified" and therefore programmed and automated, whereas the last level, the fourth, cannot. This is so because wisdom includes ethical and moral aspects that cannot be bound by any uniform rules, because they depend on the decisions of the stakeholders who influence and implement the data transformation process but not the product. In [1] he concludes that

> *"... wisdom-generating systems are ones that man will never be able to assign to automata. It may well be that wisdom, which is essential to the effective pursuit of ideals, and the pursuit of ideals itself, are the characteristics that differentiate man from machines."*

This statement, of course, depends heavily on what one considers wisdom to be in the context of artificial intelligence. Is there a reasonable definition of wisdom that will be *computationally* realizable, and can therefore be assigned to machines? How do we view Ackoff's claim in light of the current trends in the development of artificial intelligence? We will see in the next sections.

## 3 From data to wisdom

In what follows, we consider AI systems to be presented as *embodied cognitive agents*. These agents are physical entities that can continuously perceive their environment, predict the course of events, act expediently and ethically to meet their goals, learn from their experience, and adapt to changing circumstances.

If we want to understand how such an AI system works, we must look at what data and information the system operates with, what knowledge it can generate and use, and what the "wisdom" it produces, if any, is designed to do. To understand this further, it is appropriate to look at Figure 3 again, now from the perspective of the *knowledge-based theory of computation* developed some years ago by the authors [11]. This perspective has the advantage that it works directly with the concepts as used in the definition of

the DIKW hierarchy. According to this theory, computational processes are precisely those processes that generate knowledge and wisdom from information over a given knowledge domain $D$, using a given knowledge theory $T$ for their legal inferences. Hence, returning to Figure 3, the way the transformation of the data proceeds from level to level, actually defines the concepts of the hierarchy *computationally*.

Let's see how to interpret this more concretely. A *knowledge domain* $D$ consists of the information about objects, facts, and real-world processes that are of interest to an AI system or agent. The domain elements are a subset of the real world (in a suitable representation). The knowledge domain information is supplied to the system partly from the outside and partly obtained from its input data, all read by the relevant sensors. By reading the data by a given sensor, the data becomes information of the type for which the sensor is intended. This type of data must match the type of information that constitutes the knowledge domain. The *environment* of an AI system or agent is considered to be the part of the world that is described by the knowledge domain $D$ and registered by the system's sensors.

The *control system* of an agent is a knowledge theory $T$, a more or less formal theory that captures the properties of the given knowledge domain and the ways in which new knowledge can be inferred, still within the domain. What action is considered expedient is determined by the *mission specification* of the agent, for each situation. The mission specification defines what conditions the system must obey, depending on the history of its previous actions. Note that a mission specification is not the same as a *functional specification* of the system. The former specifies what the system should do, whereas the latter specifies how it should do it.

The *computations* of the system proceed by repeatedly, or continuously, combining elements of the knowledge domain (or their representations), also called *elementary knowledge*, into derived and often more complex constructs that form new knowledge, again over the domain $D$ and within the framework of knowledge theory $T$. This processing stage corresponds to the *third* level of the DIKW hierarchy, the "knowledge" level. To combine information elements at this level, the computations use a set of (inference) rules, which are either pre-given within theory $T$ or formed through learning over a large number of different computations over the given domain.

Operationally, as soon as the system retrieves some data, this data becomes information within theory $T$. From this, the system generates knowledge in the manner described above; some of which may be further used at the *fourth* level of the hierarchy, within the *wise action generation system* (see below). At this level, another capability of the AI system comes into play, in addition to data transformation – and this is *agency*. Here, agency is the ability of a system to develop a purposeful activity toward the accomplishment of its goals or the mission for which the system was designed or has evolved.

At the fourth level, if provided, agents may have a special ability to generate and use a specific kind of knowledge for their actions - namely *(artificial) wisdom* [12]. It manifests itself in the actions (behaviors) of an agent that fulfills the mission of its system in all circumstances. *Artificial wisdom is the ability of an agent to apply its knowledge, in all circumstances, to actions that are directed towards the purposeful creation of pragmatic values as prescribed in its mission specification while adhering to ethical values.*

It is now clear why knowledge and wisdom are not at the same level in the DIKW hierarchy. Knowledge is based on the acquisition of data and information and results from computations over the domain and within its knowledge theory $T$. In contrast to this, *wisdom* brings to computation a specific form of knowledge processing - namely continuous purposeful and effective action (agency) through the synergistic effect of knowledge, cognition, and action. As such, knowledge itself represents a passive form of knowledge, whereas wisdom represents an action form of it. Wisdom cannot exist without knowledge. The ability to generate artificial wisdom takes the capabilities of an agent to a qualitatively higher level compared to systems that do not have this capability. This capability must be described in its knowledge theory. Ethics, or ethical behavior, can be described as part of a knowledge theory or as a separate knowledge theory. Thus, a *wise agent* is guaranteed to produce both pragmatic and ethical values, as specified in its mission statement, in all circumstances.

Formally defined artificial wisdom makes it possible to talk about the "wisdom" of even extremely simple cognitive systems such as automatic door openers. They are "wise" because they act in such a way that they open the door (by performing the action they create pragmatic value for the person passing by) whenever they recognize such a need (cognitive ability), and behave ethically (as long as they are constructed in such a way that they do not harm anyone and nothing else is required of them). A more complex system like an autonomous vehicle, will be "wise" if it creates the desired pragmatic and ethical values through the combined effect of using its sensors and motor units - namely, bringing its user safely to his/her destination.

In the DIKW's hierarchy, wisdom is the top level, suggesting that a "higher form" of knowledge does not exist. Even "super-intelligence" is a form of wisdom. Wisdom thus appears to be the ultimate goal of artificial intelligence. It is pertinent to note here that such a goal cannot always be achieved in finite time. This is demonstrated e.g. in mathematics, where there are infinite hierarchies of knowledge theories.

We conclude that the definition of artificial wisdom is firmly anchored in the knowledge-based theory of computation. This has important implications for our considerations in Section 2, where we correlated the observed increase in "information power" of information technologies with the steady upward trend of information and knowledge processing following the levels of the DIKW hierarchy. We have now argued that all of these levels are achievable through computation. This is an observation with far-reaching consequences. But, are (artificial) wisdom-generating system realistically feasible? In the next section, we explore an example of a class of AI systems that offer at least a glimpse of having artificial wisdom-generating potential.

## 4 Large language models: intelligence without cognition

AI systems (agents) that are general "artificial wisdom-generating" systems as described do not yet exist.[4] However, we will argue that current "Large Language Models" (LLMs), like Google's PaLM, Meta's LLaMa, and OpenAI's GPT models, give us some idea of what wisdom-generating systems might look like in the future. The models are applied in realizing so-called *generative AI*, which utilizes them for learning patterns and structures from (very many) examples of how past data has been used.

[4] i.e. at the time of writing

The underlying LLMs use this knowledge to interactively generate new content of a similar quality to the learned data in order to assist users in their activities. We will focus on text-oriented LLMs.

In order to specify text-oriented LLMs as formal AI systems, we first note that their knowledge domain is a natural language (or several of them). Next, their knowledge theory is an implicit model of the world described in the given language which they build during their learning phase. Finally, the ethical theory for them will be any knowledge theory that defines desirable behaviors which are compatible with the ideas of the system's builders. As explained in Section 3, these components should suffice for a system that implements the four levels of the DIKW hierarchy.

We claim that the LLM models have the *potential* to generate artificial wisdom, as argued in [12]. We make this claim despite the fact that the models are known to be fragile (prone to catastrophic failures), not fully reliable (capable of delivering incorrect and/or contrived information), and occasionally capable of making elementary logical errors in reasoning and/or simple computation. However, in practice the models perform surprisingly well and give the "illusion" of intelligence and purposeful behavior. The claim of generating (artificial) wisdom is at least intuitively valid.

To say more, we consider how LLMs can exploit their knowledge theory. To begin with, we will argue that the LLM models have at least potentially the capabilities of (embodied) cognitive agents. To this end we call on a paradigm that is little used in AI, cybernetics, or robotics, but all the more familiar in cognitive science: the *4E cognition* paradigm [6]. The paradigm postulates that (human) cognition is not merely an internal, individual process, but an emergent process that emerges from the interaction between the brain, the body, the environment, and the social context.

The abbreviation 4E refers to the claim that cognition is embodied, embedded, enacted, and extended, namely by processes and structures outside of the brain. *Embodiment* refers to cognition as being anchored in our senses, bodies, and physical experience. *Embeddedness* means that cognition is facilitated by our environment and our way of life. *Enactment* means that cognition is served by purposeful action in the real world. Finally, *extendedness* means that the cognitive system is seen as a whole that includes not only the brain but also the means of the environment, including other people, tools, and devices. Proponents of 4E cognition argue that the four attributes are indicative of a "thinking ability" and intelligence of the system that exhibit them.

How can this be interpreted for LLMs? Clearly the attributes of 4E cognition cannot be directly applied to LLM systems, if only because they do not have bodies, senses, or effectors. However, considering the four attributes of 4E cognition, we can at least speak of *indirect* embodiment, *indirect* embeddedness, *indirect* enactment, and *indirect* extentedness in the context of such systems. The "indirectness" stems from the fact that LLMs do not have a cognitive apparatus that would allow them to interact directly with their environment. However, as we mentioned above, their knowledge theory contains an implicit model of the world that they extracted from descriptions of the world and its properties, obtained "second-hand", mediated by people in the written materials from which the model was learned. Note that in such a case the LLMs do not perceive the world "as it is", i.e. as we humans directly perceive it, but only (or exactly) as it is written about, including by artificial intelligence.

Indirect embodiment, indirect embeddedness, indirect enactment, and indirect extensibility are not the same as the corresponding attributes considered in the classical

case of 4E cognition. It leads an agent to a behavior based on indirect rather than real, immediate cognition. (It is comparable to how humans act when they arrive at a decision based on "thinking".) However, even so, in this case, we can at least infer from the actions of the system a kind of *illusory intelligence*, which offers the illusion of intelligence relative to the given environment, based on a massive aggregation of data from that environment and a selection of responses that depend on the current and past context of the system's actions. The more on-line an LLM is and the faster it is updated, the better an agent's reactions can be expected to be.

Given the illusory intelligence of LLMs, we can also speak of the *"illusory wisdom"* that the systems produce. Note that the texts an LLM generates actually constitute *artificial wisdom*. The semantics of the texts generated by the model represent the desired pragmatic or ethical values, and the computational act of constructing "responses" to "interactions" corresponds (or should correspond) to the "purposeful application" of the model's knowledge. The "purposeful act" is query-driven - the LLM generates the text that best matches the answer to the given query.

The illusory intelligence and wisdom of LLMs tends to be quite impressive in practice. However, it also explains the susceptibility of LLMs to catastrophic failures when they generate contrived information or make elementary logical reasoning errors. This is because the inference mechanism of these systems is based on statistics, not logic. The system cannot generate "true answers" in every case because it does not have access to the outside world "directly" or to other independent sources to verify facts. The system can only generate facts that do not conflict with its learned training data. It is not yet clear whether this so-called "chatter" of large language models is an inherent feature of them that we can only minimize but not completely eliminate.

Illusory intelligence is better than no intelligence in many cases. The question is whether some form of consciousness is part of such illusory intelligence. Illusory consciousness? Considerations of whether LLMs can have consciousness are beginning to appear in the work of experts in both philosophy and artificial intelligence.

## 5 The art of asking: how to get wise answers from LLMs

What wisdom can we get, at least in principle, from large language models? This question is not an easy one to answer. First of all, it depends on what a given LLM is actually modeling. Speaking very generally, and keeping to text-oriented LLMS again, we know that LLMs are actually implicit models of (some part of) our world, captured in natural language. They are trained on huge amounts of textual data that contain information about our world - facts, information, patterns, linguistic structures and knowledge, and the relationships between them. This allows the LLMs to acquire and derive further knowledge and even (artificial) wisdom about the modeled part of the world, and to generate texts consistent with what they know about it.

As users we can get knowledge from an LLM through queries called *prompts*. The knowledge, and thus the wisdom, we can obtain appears to be very dependent on how an LLM is queried. Certainly, the more precise a query, the better formulated the request, the more accurate and to-the-point the system's response can be expected to be. In order to formulate queries effectively and be a good *prompt engineer*, it is important to know, at least in a rough outline, how LLMs work. This will allow one to make use of the characteristics of the model when constructing queries and enable it

to answer queries efficiently and accurately. We give a highly simplified explanation of how an LLM operates, free from less essential details, in order to highlight some aspect of this.

The chief data structure underlying an LLM is a huge network. The nodes store information about words and sentences. Connections between the nodes represent relationships between them that are built through learning. The network can be seen as an analogy of the brain: the nodes are like brain cells - neurons, and the connections between them correspond to synapses. In technical terms, the network is an *artificial neural network* (ANN).

The neural network is trained, or "learns", by reading texts from the Internet. New relationships between words and sentences are expressed by new connections between nodes. Each connection is assigned a weight. The weights are vectors that characterize the semantic properties of the relationships: they represent their strength, type, and properties. The vectors are built and modified during the learning process. This training, or learning phase, proceeds by absorbing as much of the knowledge and wisdom as possible from books and websites that pertain to the system's mission.

Once the neural network is trained, the system can be asked to respond to prompts. Acting on a given prompt, the system generates the response word by word. The generation of each subsequent word depends on the entire prompt, the partial response the system has generated up to the given point, and of course on what the model has learned from its training data. To this end, the model has a very important mechanism that is at the heart of its "intelligence": the *attention mechanism*.

The attention mechanism is used by the model to select, predict, and focus on the word that the model will generate next. The idea is to select the next word that will contribute to the "right" answer, the one that would best respond to the given prompt in terms of the knowledge stored in the network. In doing so, words that are irrelevant for this purpose are omitted. (These other words may, however, be important for formulating a grammatically correct response.)

The attention mechanism is, again, implemented using a neural network. For its operation, it uses the weights of the connections between nodes that express the strength (the degree of probability) of the relationship between the words of the prompt and the words of the partial answer. The attention mechanism then selects the word that is most likely to "fit" the extension of the already generated partial response and also the words from the prompt. This ensures that the system generates the semantically most appropriate and grammatically correct answer.

At the same time, the system monitors the whole process by tracking the degree of similarity (again using weights) between the prompt and the answer generated so far. If this degree of similarity starts to decrease, the system "knows" that it is not generating the most relevant response and triggers the attention mechanism to focus attention on other parts of the prompt. This process of monitoring the degree of similarity and using the attention mechanism is repeated until a complete and meaningful matching response is generated.

Monitoring the degree of similarity and exploiting the attentional mechanism is one of the fundamental innovations behind the success of current generative AI. The effect is that a model does not need to access all the training data it has learned from to search for relevant information in its huge database. An LLM system works with the trained model's knowledge of the data and the discovered relationships between the

data. The language model is much smaller than its training data. The principle of LLM systems has been known for many years and is the result of a variety of mathematical and engineering inventions. However, the fact that it only starts to work brilliantly at scale, at the level of massive data and with the deployment of enormous computing power, is considered by some to be literally a *scientific discovery*.

From the principle of operation of the models as just described, it is clear that a user can influence the quality of his/her answers in only one way - and that is the formulation of the prompt. The process of refining the wording of a prompt in order to obtain the most concise and precise response, also called "*prompt engineering*", can also be understood as "hint construction" or "hinting". The purpose of a hint is to provide the LLM system - more precisely, its attention mechanism - with a concise and succinct expression that will enable the model to better understand the prompt and generate a comprehensive response. Therefore, the language of the prompt should be unambiguous and sober to achieve a good and objective response.

It often helps to mention terms in a prompt that we expect to be present in the answer. Intuition, emotion, and subjective experience are other matters that can be included if these are relevant factors for obtaining an answer. One could also mention ethical dilemmas one would like to resolve in the context of the prompt, social aspects, etc. - in short, any circumstances that the model might need to pay attention to in the framework of its artificial wisdom. This will cause the model to focus on these aspects as well when constructing a response. It is also a good idea to specify the format, form, and style of the answer - bulleted lists, tables, reflections, polite requests, popular or technical text, alternatives, etcetera.

The best way to learn how to formulate effective prompts to achieve accurate and relevant answers is not to be afraid to experiment with different forms of prompts. For certain types of questions, a well-formulated question often is a guide to its answer. Finally, it may be wise to query different models with the same content focus and observe the overlap, or the differences, between the answers.

## 6 Can we trust artificial intelligence?

What if an AI system (or agent) *deceives* us, intentionally - if it was designed for this purpose, or unintentionally - if there is a flaw in its design, its training data or in its knowledge theory? In both cases, of course, great harm can be done without being aware of it, until it is possibly too late to turn back the consequences. Some people - and often AI experts - warn that the unguided development of AI systems is prone to such risks and could be so dangerous that it may threaten the survival of human civilization. Is this true? On what arguments are such claims based?

The following result from the philosophy of computer science comes to mind. Omitting details, the result essentially tells us that, except in trivial cases,

> *there is no general effective procedure (algorithm) to decide either empirically, by testing its behavior, or theoretically, by knowing its description, whether a given AI system will be safe under all circumstances – i.e., will always act in accordance with specified "human values".*

The trivial cases we excluded are those in which we know that the specified values are either respected by all systems or respected by none. In all other cases, and these

will be common for the values and systems we consider, the result is highly meaningful. The reason in the first case is straightforward: an intelligent system, if tested, could just pretend to be a "good boy" during the (finite) test phase, but behave any way it pleases afterwards. In the second case, the proof is based on a profound result from computability theory which goes back to Alan M. Turing, the founder of modern computer science and noted visionary of artificial intelligence.

Although we have skipped over various modeling details and assumptions, the result strongly suggests that there is no general 'procedure' for deciding whether a given AI system will threaten the survival of our civilization or any other specifiable, non-trivial human value for that matter. Thinking of deciding "safety" by testing seems hopeless at first sight, and therefore it seems reasonable not to embark on the design of such systems!

However, let us look at the previous statement in more detail. It does not say that for *specific* AI systems, there can be no proof that the system will be safe under all circumstances. In other words, it may well be the case that for a specific AI system there is "clear and convincing evidence" that it is trustworthy, evidence that will only "work" for this one system. Whether we can find such evidence depends, apart from our intelligence, on the complexity of the system in question. For example, for the case of the automatic door mentioned above, it is probably possible, and for a self-driving car probably also. Hence, safety may well be feasible in many practical cases.

But what about more complex systems, such as LLM systems, not to mention superintelligent systems? Here another obstacle stands in the way, and that is the potential *opacity* of such systems. Since these systems have often developed over time through learning, a stochastic and evolutionary process, we may not even know anymore how or why they work in particular cases. They are (or, have become) "black boxes" for us. Because of the astronomical number of situations in which these systems are supposed to operate, it is impossible to specify their behavior in all the circumstances that may arise. Thus, it may not even be possible to prescribe how they should behave in all circumstances, due to the multitude and non-uniformity of cases. This again seems to be a strong argument against the development of general artificial intelligence systems (AGIs), or superintelligent systems.

The hopelessness of the task of "taming" sophisticated AI systems seems even more obvious from the following, philosophical, point of view. We, humans, having a certain intelligence, are faced with the task of finding a mechanism to prevent, under any circumstance, a "far superior intelligence" from escaping our interests and control. Is it at all possible, in principle, for a "lower intelligence" to direct, as it were, remotely, in the future, the behavior of a "higher intelligence" in this sense? Ilya Sutskever, co-founder of OpenAI, draws a parallel with the case of babies. Their parents, far more intelligent entities than infants, care very intensively for their offspring. Hence it is somehow possible to "imprint" this duty, to care for the well-being of the lower intelligence, into the functioning of the higher intelligence.

Our apparent inability to detect the mechanism behind such behavior is a huge obstacle on the road to universal artificial intelligence. To mitigate fears of an existential threat to humanity from artificial intelligence, it is necessary to realize that current artificial intelligence is far from threatening us existentially. It simply cannot, because it can at most control the minds of some gullible humans, but no doomsday machines. In the end, it doesn't even decide anything on its own. . . so far.

Instead of worrying about AI, we need to focus our efforts on explaining the real benefits of AI, in relation to our well-being, to report truthfully on AI research efforts oriented towards solving the problem of AI's compatibility with human interests, and on developing AI systems that are safe by design. It is quite possible that one day, in the future, such a task will be solved by AI systems themselves. Under these conditions, there is all reason for continuing to research, develop, and use such systems.

## 7 How it all works out

Let us now return to the two leading questions from the beginning of this essay. The questions are at the basis of the debate over the development of advanced AI systems, in particular of generative AI, and are raised repeatedly. What conclusions does our analysis lead to?

The reasons why we develop AI are often based on examples of how AI can be used as a *technology* that can enhance, replace or scale tasks that are typically assumed to require human intervention or intelligence: analyzing large datasets, faster and better decision making, automation by means of intelligent systems, up to mimicking human actions in complex situations, in a great variety of industrial or societal contexts. The applications often motivate their development already by themselves, achieving quality levels that cannot be matched or reached by humans.

For more intrinsic arguments however, we have to look deeper. What are the motives for the development of AI in general? Again this may be considered from different vantage points. Is it all about understanding human cognition and the creation of intelligent machinery? Is this a sufficient reason why, for example, the development of generative AI should be continued? Manyika *et al.*[4] argue that, from their perspective, AI-based tools, products, and services are being developed "to benefit people and society, to assist and improve the lives of many". How does this all add up?

Considering the trend analysis and arguments brought forward in the previous sections, we are led to the following answers promised at beginning of our essay:

> *We develop artificial intelligence in order to create and develop tools for generating artificial wisdom.*
>
> *The purpose of using artificial intelligence is to generate wise decisions and wise behavior through collaboration between humans and automated agents.*

In practical terms it means that the desire of *"wise" cyber-physical AI systems* and the prospects of their (wise) deployment are the primary reason why we develop artificial intelligence. It also answers the question why generative AI is being developed, or should be.

It continues to be important that the systems are kept aligned with our values. We need to look for ways to develop the systems, to live with them, and to use them safely. One solution obviously is: *human-AI cooperation*, the cooperation of humans with AI systems. The idea is to solve problems together, in mutual agreement, and not to allow one party to make major decisions without consulting the other. The development of the systems should be "batched" as in the modern theory of *phased cognitive development*, which sees the nature and development of (human) intelligence as a gradual expansion of its knowledge domains. And one should not develop the next, "higher

numbered" version of a wise system unless it is clear that the current version is consistent with our values.

The idea of human-AI cooperation is well captured by the *parental metaphor*: the relationship between humans and AI systems should be like the relationship between parents and their children. When children are young, they are nurtured and cared for by their parents; when children are grown and parents get older, they are cared for and cherished by their offspring. In this way, human and artificial intelligence will become an unbeatable duo that will allow us to shape our future as we wish.

It should be remembered that we must apply the same criteria to the wisdom generated by human-AI cooperation as applied to artificial wisdom in Section 3, namely, that it should be aimed at applying all knowledge to actions aimed at the purposeful creation of pragmatic values to cultivate our common future while respecting our ethical values. From a methodological perspective, the aim of artificial wisdom represents a qualitative shift of the computational paradigm in artificial intelligence, namely from viewing computation as a knowledge-generating process to one that generates and uses wisdom. The concept of artificial wisdom, alongside human wisdom, is thus becoming a new, and arguably the final, frontier of human endeavor.

Simply stating that generative AI should be aimed at "wise systems" does not mean or imply that its development is always or automatically targeted as proclaimed. One cannot abandon the ongoing efforts to *regulate* new AI development, such as expressed in the EU's *AI Act* [3] and pursued by many other intitutions elsewhere [2, 5, 7–9]. The *United Nations*, in fact, advocates the need for "globally coordinated AI governance", to "harness AI for humanity" [10]. In the presence of such regulations, the collaborative development of intelligent systems is a major challenge, but the benefits will be worth it. They will represent a huge advance never seen before in human history, surpassing the Industrial Revolution in impact.

Deciding to develop and use artificial intelligence tools to generate artificial wisdom requires courage, wisdom, knowledge of the issues, and faith in the ability of humans to work with artificial intelligence. At the same time, artificial wisdom and natural wisdom will constitute a lasting and meaningful legacy to our contemporaries and descendants and will increase the likelihood and quality of our survival in changing future times and unknown places in the Universe.

Wise systems will enable a better and more responsible judgment of developments in many different areas, to make better decisions, to help solve global and other practical problems, and ultimately to live more meaningful lives. It is the pathway to our future mentioned in the title of this essay. The knowledge society will (ideally…) be transformed into a wise society, a *societas sapientiae*, characterized by the purposeful use of wisdom.

## 8 Conclusion

Wisdom-generating artificial intelligence systems are fundamentally new objects that humanity has not yet encountered[5]. Their advent marks a milestone in the history of mankind because it brings a new and unprecedented form of intelligence that will be both artificial and wise and will be both complementary and alternative to our natural

[5] i.e., at the time of writing.

intelligence. These systems will understand the objective world far better than we do, thanks to their additional extra-human senses, learn faster, and be more creative.

Does this mean the end of Ackoff's prediction (cf. Section 2) that "wisdom generating systems are ones that man will never be able to assign to automata"? If wisdom is strictly meant to be *human* wisdom, then perhaps Ackoff is right: machines themselves will probably never have the qualities of empathy, intuition, or conscious experience that humans have. However, our reasoning suggests that *artificial* wisdom, given its universality and the breadth of the problem space that it can potentially cover, will surpass human wisdom in most areas.

The resulting systems will reason and argue much like humans, but thanks to their enhanced cognitive abilities, they will rely on data, information, knowledge, and wisdom that will never be accessible to humans without these technologies. They will literally become the *fountain of all wisdom*, the *philosopher's stone*, and *the holy grail of humanity*.

However, this dream will have its downsides. Unless we can satisfactorily resolve the problem of the alignment of wisdom-generating systems with human values, we will lose our dominant position among intelligent entities without their cooperation. We will be confronted with a *Faustian dilemma* – a potential trade-off between the benefits of artificial wisdom and the dangers associated with its development and use. Which path will we choose? How will we decide?

Are we ready to decide?

**Acknowledgment** The research of the first author was partially supported by the Czech Technological Agency grant EBAVEL, no. CK04000150, Programme Strategy AV21 "Philosophy and Artificial Intelligence" of the Czech Academy of Sciences, and the Karel Čapek Center for Values in Science and Technology.

## Literature

1. Ackoff, R.L., From Data to Wisdom – Presidential Address to ISGSR, June 1988, *J. Applied Systems Analysis* 18 (1989) 3-9
2. ACM Technology Policy Council, *Principles for the Development, Deployment, and Use of Generative AI Technologies,* June 2023, https://www.acm.org/binaries/content/assets/public-policy/ustpc-approved-generative-ai-principles
3. EU, *AI Act*, 2023, https://www.artificial-intelligence-act.com/
4. Manyika,J., Dean, J., Hassabis, D., Croak, M., Picha, S., Why we focus on AI (and to what end), Google AI, January 2023, https://ai.google/static/documents/google-why-we-focus-on-ai.pdf
5. Lorenz, P., Perset, K., Berryhill, J, *Initial policy considerations for generative artificial intelligence*, OECD Artificial Intelligence Papers, No. 1, OECD Publishing, Paris, https://doi.org/10.1787/fae2d1e6-en.
6. Newen, A., De Bruin, L., & Gallagher, S. (eds), *The Oxford Handbook of 4E Cognition,* Oxford Library of Psychology (2018; online edn, Oxford Academic, 9 Oct. 2018)
7. OECD, *AI language models: Technological, socio-Economic and policy considerations*, OECD Digital Economy Papers, No. 352, OECD Publishing, Paris, 2023, https://doi.org/10.1787/13d38f92-en
8. Office of the Privacy Commissioner of Canada, *Principles for responsible, trustworthy and privacy-protective generative AI technologies*, December 2023, https://www.priv.gc.ca/en/privacy-topics/technology/artificial-intelligence/gd_principles_ai/
9. The White House, *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, October 2023, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

10. United Nations, *Governing AI for Humanity*, Interim Report, December 2023, https://www.un.org/techenvoy/sites/www.un.org.techenvoy/files/ai_advisory_body_interim_report.pdf
11. Wiedermann, J., van Leeuwen, J., What is Computation: An Epistemic Approach, in: G.F. Italiano *et al.* (eds), *SOFSEM 2015: Theory and Practice of Computer Science*, Lecture Notes in Computer Science, vol 8939, Springer, Berlin, 2015, pp 1-13, https://doi.org/10.1007/978-3-662-46078-8_1
12. Wiedermann, J., van Leeuwen, J., From Knowledge to Wisdom: The Power of Large Language Models in AI, *Technical Report* UU-PCS-2023-01, Utrecht University, July 2023