Risk Machine? Risk Human? Can AI Help?
A study from the perspective of the philosophy of science

Gerd Doeben-Henisch

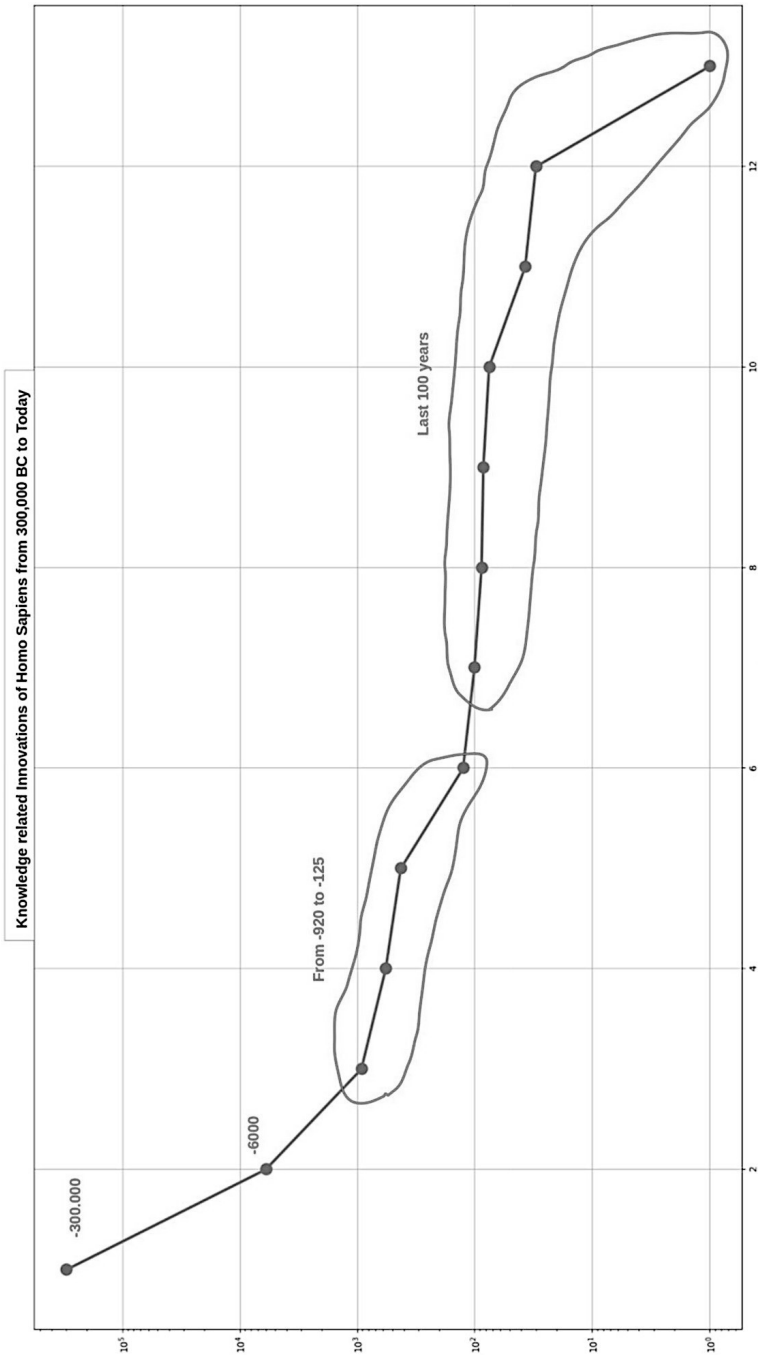
In a turbulent development of empirical science, democracy, sustainability, digitalization, and now also artificial intelligence, ever larger spheres of action have opened up. At the same time, risks are becoming overwhelmingly visible: Will global problems overwhelm us? Are we humans the problem? Can AI help here, or is ultimately the development of AI itself a new problem? The following sketch of the current situation tries to work out that the various forms of risks cannot be divided from each other. They are all interconnected. It will be crucial to shape this profound interconnection.

A. A Simple Timeline

For the sketch presented here, some historical data are provided in advance, suggesting a connection that seems important for understanding the current challenges. Looking back from the year 2024 in our calendar, there have been cells on our planet for about 3.8 billion years that indicate the beginning of biological life. However, traces of the life form to which we humans belong, Homo sapiens, only appear from about 300,000 years ago. This is a point in time that occurs after 99.99% of the preceding time. Evolutionary biology can tell us a lot about what happened in the time before Homo sapiens. Here, it only counts that we have been actors on this planet only since this relatively short time. And it is only about 6,000 years since we humans invented and used writing systems to improve our communication. This happened after about 98% of the time since the appearance of Homo sapiens. University forms of education can be observed for about 920 years, i.e., after 99.7% of the time. We have known modern printing for about 570 years (after 99.8% of the time). Modern empirical sciences began about 425 years ago (after 99.85%). Modern formal logic and mathematics have been found at least since about 125 years ago (after 99.95%). Modern democracies since about 100 years ago (after 99.96%). The concept of the universal Turing machine has existed for 87 years (after 99.97%), soon

followed by the first ideas of system engineering (after 99.97%) and about 75 years ago with the first ideas for artificial intelligence (after 99.97%), albeit largely only among specialists. The first comprehensive idea of a sustainable society – here the Brundtland Report – was created 37 years ago. The Internet, as we know it as the World-Wide Web (WWW), has existed for 30 years (after 99.987%), and a generative artificial intelligence that has made it into the everyday lives of many people, including those who are not computer scientists, has existed for 19 months (after 99.999%).

Innovative Event	*Label*	*Time BC*
Homo sapiens	1	300.000
Writing	2	6.000
University	3	920
Printing Press	4	570
Empirical Science	5	425
Mathematics and Formal Logic	6	125
Democracies	7	100
Turing Machine	8	87
Systems Engineering	9	84
AI	10	75
Sustainability	11	37
World Wide Web	12	30
Generative AI	13	1



These data in themselves may not mean anything. However, the sometimes enormous spans of time between individual events can indicate the tremendous complexity that had to be managed in the ongoing development. It is striking that significant achievements of Homo sapiens have all occurred in the last 2% of its sojourn on planet Earth, with a particular concentration in the last 0.3%. Mathematically, this can also be seen as a form of 'increasing event density,' also as a kind of 'acceleration': more and more in less time and simultaneously with an increase in complexity.

B. Cluster Effects

Upon closer examination, it also becomes apparent that these events are not independent of each other. The increasingly complex scientific and cultural achievements of humans over time are not possible without intensive and efficient coordination of many individual brains with each other. Without communication, this would be impossible. This requires suitable languages and sign systems, as well as highly networked work methods that rely on shared knowledge (books, journals, libraries, databases, internet, ...). Information-rich and verifiable linguistic communication is a minimum (standardizations, empiricism and prognosis, mathematics, ...). The increasing liberation from time and place (internet, databases, mobile networks, distributed data collection, ...) is added, as well as the management of ever larger amounts of data and the automation of routine tasks (algorithms, computers).

This short list already shows that the many new techniques and technologies did not arise 'just like that.' They were triggered by corresponding demand, and they were possible because the 'collective thinking of humans' was capable of ever greater achievements and continues to be so.

C. Irritations

When considering how many paths biological life on Earth had to take over 3.8 billion years to keep life on this planet 'in the game,' it should not be surprising that there can also be challenges in the current phase of life on the planet that can somewhat disrupt the 'usual course of business of the last centuries or even millennia.' The mere occurrence of such events perceived as 'disturbances' does not necessarily mean that the project of

life on the planet is fundamentally in question (there have been many events in the past that, even from today's perspective, appear so enormously threatening that the present may seem harmless by comparison).

For simplification, the current challenges for the following discussion will be grouped as follows:

1. Changes in the Earth system that threaten the existing habitats and habits of humans as well as large parts of the entire ecosystem.
2. Changes in the everyday structures of human societies, which can have various reasons, here those that have nothing to do with digitalization.
3. Changes in the everyday structures of human societies that are related to digitalization.

In the following, the changes in everyday life in the sense of point (3) will be considered, i.e., those related to digitalization.

D. Irritations in the Context of Digitalization

While the new methods of system engineering in conjunction with digitalization, modern empirical science capable of prediction, and increasingly intelligent programs have proven to be enormously powerful and continue to prove so daily, there are constellations in societies that call this fundamental capability into question. Here are some examples:

- The focus of the development and deployment of this new socially relevant technology cluster is predominantly in the hands of private companies, whose interests are not the same as those of the overall society. Important further developments may thus be blocked.
- Large parts of the users of the new technologies largely lack a sufficient understanding of the effects of the system on themselves as individuals and on entire user groups.
- For democracies, a functioning common public sphere is vital. For years, we have been experiencing a fragmentation of one potential public sphere into many 'quasi-private' public spheres (in some cases up to 90% of a country's population) due to the comprehensive availability of the internet, which not only prevents the formation of a sufficiently common opinion but also accompanies this fragmentation with streams of opinions that promote the formation of enemy images among each other and 'false truths.'

– The successful deployment of digital technologies pays off economically. This reinforces further development, where 'intelligent programs' have a sales value that gives them significant importance in the thinking and feeling of people. In direct comparison with the collective power of humans, these programs are rather simple, but the interest in the collective intelligence of humans is thereby factually weakened. This is not without danger.

In the following, points (1) - (3) will not be discussed further, although they may be of high societal relevance. Instead, further thought will be given to point (4), namely the relationship between 'Collective Human Intelligence' and 'Artificial Intelligence'.

E. Paradox: The Disappearance of Genius

A paradox: In earlier times, when it was individual people who produced outstanding achievements (painters, architects, war heroes, captains, musicians, composers, ...), these individuals enjoyed high and highest esteem, and people were even willing to see in them a 'genius' at work, a 'divine spark,' the 'world spirit,' and similar concepts. However, as the actions spread across more and more people, large workshops, networks, complex working groups with many thousands of experts with different focuses, the 'human genius' became less and less tangible. When thousands of scientists, engineers, and various workers create great structures, bridges, rockets, airplanes, ships, one still sees the product, may still be impressed, but the collective human achievement behind it becomes strangely invisible, disappears. An ordinary individual is usually no longer capable of even remotely grasping the entire collective effort behind it. How could they? When 10,000 or more people research and work together in highly complex ways for years, who can understand this process? To whom is it attributable, and who bears the responsibility? Schools often still operate in the realm of old work and knowledge models that no longer exist, and even a university is far from these fantastic achievements of modern engineering; and where do modern media stand? Newspapers, television, podcasts... one can search for a long time and find virtually nothing about the reality of modern collective intelligence. Are we humans making ourselves invisible?

F. Intelligence in Humans and Machines

I. Measuring Intelligence

In the context of modern psychology, there has been the concept of intelligence as the 'Intelligence Quotient' (IQ) since at least the beginning of the 20th century. This does not mean that one knows what 'intelligence' is, but one knows how to measure certain behavioural performances of people in such a way that the measurement result can be labelled 'Intelligence Quotient' (IQ).

This concept of intelligence was based on the assumption that a list of typical tasks that a group of people of the same age in a certain region can usually solve provides an indication of which behaviours should be called 'intelligent' (what other reference point should one have chosen?). In the evaluation, one obtains the number of correctly solved tasks for each person. Assuming a 'normal distribution (Gaussian distribution)' of the values, one can set the mean value as 100 (IQ of 100), and arrange the weaker or stronger values 'left and right.' For the chosen tasks and the chosen group, one can then assign an IQ value from the distribution to each person. If one disregards all the nuances and conditions, then the IQ value here functions as an index related to a set of selected tasks and the correlating observable behavior of the acting agents. Of course, this implies nothing about the 'internal structure' of an actor that may be available in the actor and responsible for whether and how an actor behaves.[1]

Considering the range of possible human behavior in relation to the many action situations that are possible in everyday life, the typical collections of tasks already seem somewhat overly simplistic, especially when one knows how great the individual variability of characteristics is and that in real life it is not only about the individual behavioural characteristics in isolation, but also and increasingly about the ability to solve difficult tasks 'in the collective with others.' This requires many abilities that are hardly of significance for an individual. These considerations are not meant to deny that individual tests can nevertheless provide indications of a person's

1 In the history of IQ measurement, there have been collections of tasks that have challenged a wide variety of abilities, and there have been attempts to correlate the behavioral data with 'hypothetical structures inside the actor.' However, none of these approaches have been entirely convincing to date. So far, no integrating model has been presented that can uniformly process all these different sets of tasks.

performance potential, but these indications should be critically placed in larger contexts.

II. Cognition and Intelligence

At the end of the 19th century, modern psychology dealt with many cognitive performances of humans independently of intelligence tests. These included topics such as 'perception,' 'memory,' 'language learning,' 'language understanding,' and much more. These research works were based on targeted experiments, which then formed the starting point to develop hypotheses about functional structures 'in the actor.' All the hypotheses together can then function as a 'functional model' designed to derive 'predictions of behavior' from it.

If the 'models of cognition' were suitable for processing tasks from intelligence tests, such models could also be correlated with an IQ value. In this way, cognitive models of humans could then be indirectly evaluated using IQ tests.

III. Intelligence in Psychology and AI

Although Alan M. Turing had openly contemplated the possibilities of machine intelligence as early as 1948 and discussions about intelligence and 'intelligent machines' in computer science became a constant companion, the way computer science deals with intelligence and the way psychology does it have never really converged. Computer science has always had a strongly pragmatic approach and examined the capabilities of algorithms in specific task scenarios. Generalization has been and remains difficult with this approach.

IV. Human and AI

From these preliminary remarks, it becomes clear that a unified discussion about intelligence in humans and machines is currently still difficult. Unified task scenarios would be a first step. In addition, increasing the diversity of scenarios. This would at least make it possible to provide a rough estimate of the strengths and weaknesses of the two types of actors (human and machine).

Currently, the discussion about the relationship between machine and human intelligence is very unsatisfactory, especially since the possibly most important aspect of human intelligence, the so-called 'Collective Human Intelligence,' is still rather 'unexplored'.

V. Collective Human Intelligence and AI

Research on collective human intelligence is overall not very advanced yet[2], but there are already new works on the topic of 'Hybrid Collective Intelligence,' which investigates the interplay between collective human intelligence and machine intelligence.[3]

Here, as an example of collective human intelligence, a modern application scenario is taken, which is prototypical for collective intelligence: a development process in the style of system engineering. In this, the role of collective human intelligence can be made visible, including how it can generate machine intelligence. In this context, further open points can be clarified.

VI. Human in System Engineering

As already noted in the introduction, the tasks of the modern age require not just the efforts of individual masters and their assistants but huge teams, often with many thousands of experts, possibly distributed across many locations. Without efficient communication underpinned by corresponding documents, a valid result is out of the question. In addition, numerous abilities are necessary beyond mere cognition for human collaboration to

2 For example, see the MIT project 'Handbook of Collective Intelligence,' edited by Thomas W. Malone and Michael S. Bernstein, URL: https://cci.mit.edu/cichapterlinks/ or 'Understanding Collective Intelligence: Investigating the Role of Collective Memory, Attention, and Reasoning Processes' by Anita Williams Woolley and Pranav Gupta, URL: https://kilthub.cmu.edu/articles/journal_contribution/Understanding_Collective_Intelligence_Investigating_the_Role_of_Collective_Memory_Attention_and_Reasoning_Processes/24049830/1.

3 For example, see 'Collective Intelligence in Human-AI Teams: A Bayesian Theory of Mind Approach' by Samuel West and Christoph Riedl, Proceedings of the 37th AAAI Conference on Artificial Intelligence (2023), August 24, 2022, URL: https://www.networkscienceinstitute.org/publications/collective-intelligence-in-human-ai-teams-a-bayesian-theory-of-mind-approach.

function over the long term and even under stress. An orderly creation process must be organized, in which all human actors work together communicatively coordinated from an initial idea to a real product or a real service. One type of such joint creation processes is called 'System Engineering,' and the whole process is the 'System Engineering Process (SEP).'[4]

So, simplifying, there is the group of human experts EXP_{HS}, who both create the important documents DX and then use these documents as guidelines for the implementation of an order. Specifically, simplifying, the documents are:

1. Problem statement $D_{problem}$: Description of which problem is to be solved.
2. Requirements D_{requ}: Translation of the problem statement into concrete requirements.
3. Technical Design Document D_{design}: Translation of the requirements into concrete technical design decisions.

Important at this point is that all documents consist of the character strings (STR) of a particular language L (STRL). These character strings have an associated meaning space MEAN(STRL), which itself is not present as a document but exists exclusively in the form of 'internal states (IS)' within an acting agent. This highlights a special characteristic of the human actors in this process. They have the ability to link the character strings of a language L with internal knowledge states IS_{know} so that all participants in the language can activate these internal knowledge states through the character strings, and vice versa for internal knowledge states, they have the corresponding means of expression.

This duality of character strings STRL of a language L on one hand and internal meaning structures IS_{know} on the other, linked via a meaning assignment $MEAN_L$: STR_L <---> I_{Sknow}, enables great flexibility in constructing different meaning structures and their encoding through meaning assignments using character strings.

However, this flexibility has its price: all users of a language L must not only coordinate their interindividual knowledge contents I_{Sknow} with the

4 A formalized example of a System Engineering Process can be found here (i) Erasmus, L. D. and Doeben-Henisch, G. 2011. A Theory of the System Engineering Process. In the 10th AFRICON Conference: Sustainable Energy & Communications Development for Africa, Livingston, Zambia, and here (ii) L. D. Erasmus and G. Doeben-Henisch, A Theory of the System Engineering Management Processes in ISEM 2011 International Conference, Sept. 2011.

perceived properties of the real external world, but also the interindividual coordination of their linguistic meaning assignments to the respective character strings. This requires a continual reassessment of these assignment and coordination processes. There is no fixed point in this process!

This structure implies 'by design' a 'false normality,' as the human actor only briefly possesses sensory perception of the real external world, interpreted through prior knowledge. The 'current' then partially transitions after a brief 'moment' into the mode of the 'memorable.' 'Presence' then exists primarily in the mode of a 'memorable present.' This can—as is known—be distorted or even false. Whoever does not continually work against this distortion lives, by doing nothing, in a 'distorted world' where much is not as it is 'in the real world out there'.

The various documents $D_{problem}$, D_{requ}, and D_{design} thus do not necessarily describe the 'world as it is,' but the 'world as seen by the authors of the texts.' This is, of course, true in a very explicit way for all 'future situations.' The consequences can be varied. Since a design document D_{design} can only approximate the object M_{tst} to be realized or the service to be realized in the mode of the meaning knowledge of the authors, it may be that properties come into play in the real implementation of the linguistic concepts from the design document that are due to changed meaning spaces of the involved authors or implementers. Even if a verification of the test object M_{tst} with the design document D_{design} appears formally correct (as verification), the verification may still lead astray, as the real-world reference of the design documents may lead to conflicts due to incorrect assumptions about the world.[55] Such a 'fundamental error' can remain undetected in the context of verification, but if one begins to evaluate a test system M_{tst} with real application situations, it can happen that the assumptions about

5 This problem has long been known in the context of research on Safety-Critical Systems (SCS). For example, see Nancy G. Leveson, who has identified this problem as a fundamental issue in numerous articles and books, most recently in 2020 with N.G. Leveson. 'Are you sure your software will not kill anyone?' Communications of the ACM, 63:25 – 28, [https://doi.org/10.1145/3376127], and in 2023 with Nancy G. Leveson and John P. Thomas, 'Inside Risks Certification of Safety-Critical Systems. Seeking new approaches toward ensuring the safety of software-intensive systems.' COMMUNICATIONS OF THE ACM, OCTOBER 2023, VOL. 66, NO. 10, pp.22-26, [https://dx.doi.org/10.1145/3615860]. Also see Gerd Doeben-Henisch, 'Review of Nancy Leveson (2020), Are you sure your software will not kill anyone?' URL: [https://www.uffmm.org/2023/10/21/review-of-nancy-leveson-2020-are-you-sure-your-software-will-not-kill-anyone/] and text: [https://www.uffmm.org/wp-content/uploads/2019/06/review-leveson-2020-acm-yourSWwillNotKill.pdf].

the application situation lead to concrete conflicts when confronted with real application situations. This is the only way to discover implicit false assumptions about the real application situation.

With all this, it becomes clear that a design process with final verification and evaluation can ultimately be understood as a 'dialogue' between the previous expectations of the world and the way the real world 'actually shows' itself.

In summary, one can say: the human in the 'mode of his collective intelligence' uses the maximum of his current knowledge, which can be partially wrong due to the nature of human cognition, to generate possible new products or behaviours in a possible imagined (predicted) future. To avoid falling victim to the existing—albeit unconscious—knowledge errors, collective intelligence tries to make these visible and eliminates them during the development process through the most informative tests possible. However, this can only ever work to a limited extent, as the entire collective knowledge lags behind the surrounding dynamic complexity at any given time. Therefore, the collective knowledge must be repeatedly not only 'partially corrected' but also fundamentally adjusted.

VII. Can AI Help?

At this point, the question will be addressed whether the new forms of Artificial Intelligence (AI) can help human Collective Intelligence in any way, or whether AI could perhaps completely replace the role of collective human intelligence eventually?

Starting with the fundamental question of a possible complete replacement, this question is quickly answered, as so far neither the concept of 'collective human intelligence (CHI)' has been defined in a way that allows for verifiable comprehensive tests[6], nor does a similar definition exist for the term 'artificial intelligence (AI).' In the case of AI, there are collections of various performance tests, but it is not clear how these can be 'generalized.' Even less clear is how a connection to 'collective human intelligence' can be established as long as this term is not really defined.

6 For example, see the draft by Thomas W. Malone and Michael S. Bernstein in 'Chapter 1. Introduction' for the planned book by Thomas W. Malone & Michael S. Bernstein (Eds.), 'Collective Intelligence Handbook,' MIT Press, in press. URL: [https://docs.google.com/document/d/1CRVN8uxa_g8i3oLRfVxhsltWNZ_ZMwoI-pl5IosG9VU/edit?pli=1].

Therefore, the following will only address whether the current AI—and its possible extrapolated extensions—could help collective human intelligence in any way.

Based on the previous discussion, there are so far only two possible starting points for a discussion: Firstly, the connection of the term 'intelligence' via the 'intelligence quotient' with observable performance in the face of tasks to be solved, and secondly, certain formats in which people collectively solve tasks to which the property of 'intelligence' is assigned—rather intuitively. One such collective format is the previously mentioned 'System Engineering Process (SEP).'

Since known intelligence tests always only consider individual persons, they are methodologically of little help for the discussion of the phenomenon 'collective human intelligence.' In addition, there are so far no systematic comparisons between humans and intelligent machines for measuring individual performances.[7] Therefore, an attempt will be made here, at least for the example of a System Engineering Process, to clarify whether there are areas in which intelligent machines could support or even replace humans, or not.

G. World Models: Open or closed

I. Language: With and without meaning

In the context of a System Engineering Process, collective human intelligence (CHI) starts with a problem statement $D_{problem}$, which uses the currently available knowledge about the 'application situation' and the 'available solutions' to work out a 'concretization,' initially 'mentally' (D_{requ}, D_{design}), but then also 'materialized' with a verifiable real test version M_{tst}. This test version is then tested in a variety of 'application situations' ANW_{tst} to see if all the anticipated behavioural properties from the requirements and design document (D_{requ}, D_{design}) can be positively fulfilled.

The entirety of the agreed-upon documents ($D_{problem}$, D_{requ}, D_{design}) represents the current 'world model (WM)' of the CHI.

7 Of course, one can cite examples where individual humans have competed against computers in defined games (checkers, chess, Go, and many more), and now almost exclusively lose against computers. These examples are certainly informative, but they do not replace a real comparison with a variety of tasks.

As previously made clear, such a world model is fundamentally incomplete and highly likely to be partially inaccurate. It is a world model that is 'closed on paper,' but in the face of confrontation with the real world through real tests, it must be classified as partially changeable. From this perspective, the world model is 'partially open.' An experienced CHI 'knows this' and therefore organizes appropriate tests for verification.

The minimal elements of a world model are (i) a defined initial situation (S), (ii) a set of possible change rules (R), (iii) an agreed procedure on how to change an initial situation—or any situation—using change rules (|--), and (iv) at least one goal (G) that can serve as a benchmark to assess whether—and if so, to what extent—a current situation already corresponds to the agreed goal.

In the case of a CHI, there is also the ability (v) to 'decide' whether a currently reached linguistically defined state S 'in the light of the active meaning functions of all participants' is 'true' or not in the real situation. This would not be a purely formal verification as can be performed within a SEP, but an empirical verification that is only possible by explicit reference to the surrounding empirical world.

It is also known that a CHI is capable, in the event of conflicts between the current world model and the real world experienced in testing, of modifying its world model to the extent that the conflict no longer occurs. In the worst case, the world model would have to be 'discarded.'

Modifying a world model is not trivial. Many change rules have effects both 'in breadth' (side effects) and over many successive time points. Identifying the decisive misalignment is not easy to achieve. Additionally, complex meaning functions are interposed between the character strings of the documents and the possible reality, which can differ among the individual members of a CHI without these differences being directly visible. If the specific elements of a misalignment are discovered, the challenge arises of finding alternative change rules, possibly also a change of goal. For these creative tasks, there is often/mostly no 'rational assurance' through existing knowledge, as it often involves 'truly new' situations that no one really knows yet.

At this point, considering those intelligent algorithms that now routinely defeat the world's best players in defined game contexts, one might wonder whether this type of algorithm—let's call it a 'closed world model (CWM) algorithm'—would be suitable for making a constructive contribution in the context of a System Engineering Process (SEP).

If the world model of a CHI were fully formulated, it would be conceivable that a CWM algorithm could master this task.

Here, however, an immediate fundamental 'obstacle' becomes visible: Unlike the closed world models of a game, the rules of the world model of a CHI are predominantly 'not formalized,' as the rules are written in 'normal language,' which is not applicable without an explicit meaning function. This may initially be interpreted as a 'weakness,' but real practice shows that this 'weakness' is precisely the strength that makes a powerful CHI possible.[8]

A CWM algorithm could therefore only be applied if it were capable of not only processing 'meaning-free' character strings with 'hard-wired meaning objects'[9] but also of appropriately interpreting freely interpretable character strings with one of the many available meaning functions in the context of natural languages. Due to the radical 'meaninglessness' of computer languages, there is so far no indication of how this problem could be satisfactorily solved.[10]

II. Reality Check: True or false

As the example with the System Engineering Process makes clear, it is fundamentally important that the active world model of a Collective Human Intelligence (CHI) is repeatedly and extensively verified and validated

8 This was the attitude of those logicians and mathematicians who, from the end of the 19th century, advanced modern formal logic and the formalization of mathematics. They 'liberated' logic from any 'meaning,' except for some 'abstract truth values.' This enabled very elegant formal calculi, but when applying these to the 'real world,' all character strings of the formal logic languages had to be interpreted back to the real world in an extremely laborious and error-prone manner. All modern computers suffer from this fundamental 'withdrawal of meaning.' This is so far 'irreparable.'

9 The 'hard-wired meaning objects' in the case of game applications are those character strings to which fixed objects from the game are uniquely assigned within the scope of a game. A game rule refers to such objects and describes defined changes that can then be directly translated into a change on the game board.

10 Of course, computer languages are not 'completely devoid of meaning,' since they can be interpreted by the respective machine in such a way that character strings of the programming language can lead to state changes within the machine. However, there is so far 'no natural connection' between the state changes within the machine and possible meaning assignments of character strings of everyday language in the real world 'outside the machine.' This would have to be specially established in each individual case. So far, there is no known approach that could solve this problem (see also note (9)).

through tests with the 'real world.' This is possible because all human actors within a CHI have the ability not only to correlate language strings with learned meaning functions and acquired knowledge (also known as 'decoding,' 'interpreting,' or 'understanding'), but also to relate this knowledge activated through interpretation to current sensory perceptions of the real external world. Within this non-trivial process, a 'judgment' may be made that the 'activated knowledge' sufficiently matches the 'perceivable aspects' of the 'real world' or not. This 'empirical control' plays a fundamental role in assessing the current world model and for possible changes to this model. Without this, all world models would be nearly worthless.

Modern algorithms, such as the type of a 'generative AI' exemplified by chatGPT4 or similar programs, exhibit behaviours that can easily give the impression that they 'understand' the 'meaning of character strings of everyday language' as a human actor would. Indeed, this performance is extraordinary because the algorithms of the 'generative AI' type actually do not possess a meaning function that is comparable to that of a human actor.

This capability is based on two fundamental functions of a generative AI: (i) These AIs are 'fed' millions—or more—documents created by human actors, from which they independently 'extract' individual character strings with their various 'contexts with other character strings' and frequencies. This already allows for determining which character strings are commonly used with others. In a further step (ii), typical dialogue situations are identified with the help of human actors, and it is trained how character strings within such dialogue formats can be organized so that they correspond to conventional formats. This also happens without any explicit meaning function. The fact that such AI can generate long dialogues and extensive texts in a way that at first glance seems as if they were generated by a human actor is impressive and the result of excellent engineering work.

Due to a lack of a meaning function, which goes along with an absence of human-like world knowledge based on sensory input, further modified by various cognitively relevant brain processes, a generative AI can only move within the predefined paths of available texts. A current empirical reference is thus excluded unless there were an empirical segment of the world whose properties are translated into character strings of everyday language in real-time, in a way that this translation meets the requirements of a human meaning function. If there were such 'Real-time Empirical State Descriptions (RESD)' for a specific 'area,' then a generative AI could at least partially match its character strings with these RESD character strings.

Where in our world would there then be such RESD character string generators? Where would such RESD character string generators get their meaning function from? Would it ultimately be human actors themselves who translate a current situation into character strings using their own meaning function, which they then communicate to a generative AI?

When asked: 'Would you be able to judge, that a statement, which I would communicate to you, is 'true' or 'false'?' @chatGPT4 responds, 'I can help evaluate the accuracy of a statement based on known facts and information. Please share the statement, and I'll do my best to assess its truthfulness.' Yes, the current world model of a generative AI marks the space of possible utterances, which are either 'fit' or 'do not fit' relative to it. This includes the case that a generative AI has adopted documents that are 'inherently wrong.' When this fact is addressed in a dialogue with a generative AI, one gets many good suggestions on how to check the usability of documents or detect errors, but one does not get a clear statement from the AI that it itself cannot directly verify the empirical validity of a statement.

For the task of direct empirical verification, a generative AI falls short, but it still appears that a generative AI can be helpful for initial orientations.

III. Cooperation: Models of the other

As became clear from the description of a CHI using the example of a System Engineering Process, all human actors in such a process are required to have the ability to continuously and comprehensively communicate and cooperate with other actors in this process to enable a CHI.

This is a highly complex matter, the description of which is omitted here. Part of the task is that each human actor must not only have internalized parts of the common world model but also have minimal knowledge about all behavioural and communication structures. In particular, they need 'minimal models of the other' in their minds, enabling them to form useful 'expectations about the behavior' of the others.

H. Postscript

After these considerations, it should be clear that the various risks cannot simply be attributed to a single actor. In collective intelligence—whether purely human or hybrid—every individual actor is part of a larger entity

that acts and decides as a whole. Uncertainties and possible partial maladaptations are essential to the process of collective intelligence moving in a dynamic world. Here, truth can only ever be taken as a 'temporary state' that must be repeatedly achieved anew together. The price of success is called 'life,' and the price of failure may be 'extinction.' This fundamental fact has not changed after about 3.8 billion years of life on this planet.