Research Article ISSN 3069-9703

International Journal of Mathematic Exploration and Computer Education

Can We Go On Trusting AI?

Dietrich Brandt*, Luca Bischoni

RWTH Aachen University, Aachen, Germany.

***Correspondence author**
Dietrich Brandt,
RWTH Aachen University, Aachen, Germany.

Submitted: 21 Mar 2026; **Published:** 12 May 2026; **Published:** 2 Jun 2026

Citation: Brandt, D. & Bischoni, L. (2026). Can We Go On Trusting AI?. *Int J Math Expl & Comp Edu.3*(1):1-9.
DOI : https://doi.org/10.47485/3069-9703.1026

Abstract

This paper is somewhat different from how we usually publish our papers today. In a certain way, it is the working-life balance sheet which is dealing with several decades of AI and Automation research and development. Thus, the views on AI in this paper might be somewhat different from the views of many other colleagues who have been entering into this field of AI research and development more recently. Actually, our team started on such research about 1974, at the RWTH Aachen University in Germany. Around the year 1984, our team became part of the then new Department of Computer Science in Mechanical Engineering. From then onwards, we had been for several decades cooperatively involved in research and development of AI at the RWTH Aachen, and we have continued our active commitment for AI up to today.

Within a few years from the start, this department had become one of the largest and one of the leading interdisciplinary research departments in German universities – and well beyond Germany. Around the year 2010, it employed up to 50 full-time research staff. Our researchers came from engineering and natural sciences as well as from social sciences and humanities. These interdisciplinary research teams integrated senior researchers, PhD and graduate students, undergraduates, and the support staff.

This paper describes our approach to research and development in Automation and AI as it was based on the concept of interdisciplinary cooperation. The following question was developing fundamental weight for our research at Aachen, from the start of our department:

How can we create automated systems which can do things as well as ourselves or even better than ourselves, their creators?

Human-Centred Systems - Sullenberger

During the first period of our research and development in automation engineering, AI was merely some vague idea but we saw this early already the potential power of this concept beyond Automation, Neural Networks and Machine Learning. At that time, our own research was about developing first-generation automated systems for production industry, business administration and societal usage systems. For these tasks, we put our research emphasis on the concept that the human operator *remains in the loop*. This operator would be responsible for running the system particularly in cases of malfunctioning or disaster. We were the first to name this approach the *Human-Centered Systems* Approach. Today, this term is used worldwide with many different emphases.

There is one famous example to describe the problem and our understanding of *Human-Centeredness* in depth: in the 1990s years, Airbus was aiming for the fully automated aircraft. One of our PhD students at that time, Tania Hancke, committed herself to fight this Airbus automation philosophy based on our concept of human-centeredness. Her critical interventions took place at several international discussions (e.g. Hancke & Braune, 1993). They indeed contributed to the subsequent decisions of Airbus that the aircraft pilots have *remained in charge* of their highly automated systems up to today. Since those days, her criticism has remained visible internationally. The core of it is as follows: All technological systems are *Open Systems*, as a matter of principle – they are prone to chaos and break-downs either from within or from outside. It may mean that there is boredom for the human aircraft pilot during more than 99% of the time, but 1% or even much less may be chaos and catastrophe. In the end, there is only the human operator who has a chance to cope with chaos, anyway.

There exists one particular, more recent example for such chaos triggered from outside. We have analyzed it in one of our recent publications (O'Neill et al., 2020). It is the aircraft accident with US Airways Flight 1549 on January 15, 2009 when Sullenberger landed the plane on the Hudson River. Both engines had been disabled by a flock of geese: Several geese were sucked into the two engines which made them both stop within seconds. All people on board, however, survived the accident. Sullenberger was extremely experienced as an aircraft pilot, but he was also used to fly *gliders, which* means here: an airplane without engines. Therefore, he was prepared to treat his Airbus in this extreme case as a kind of *glider* when

landing. It is one example of how the human operator may also in the future be needed within the loop of such highly automated systems, corresponding to the Human-Centred System concept (Fig.1).

Figure 1: Sullenberger: Evacuation of the aircraft as it floats on the Hudson River

AI, however, is the new threat to this concept of *Human-Centred Systems*. In terms of technology, no pilot may be needed any more in the near future.

Would we trust AI so far that we were willing to fly across continents without a pilot in charge?

The RoboCup Competition – the Robotic Team

Let us look more closely at chaos created from within technical systems. Here, we may refer to the tendency of AI to deliver output which is merely a *hallucination*. The system itself seems to be ashamed of not being able to answer certain questions or to resolve tasks posed to it by humans. Thus, it starts inventing and *hallucinating* answers or data out of the blue as if to meet our expectations – or the expectations of its human creators. Obviously, taking the complexity of large AI systems into account: there is no way visible yet to retrace and correct such deep-fake system output nor such hallucinations, nor do we have the means to cut them out right away. It seems that we do not have the power at all to separate systemic mistakes and lies from truth and facts as it would be important or relevant for us – and the AI systems have been designed in ways not to grant us this power.

Very early in the development of such systems, our team was feeling confronted with such experiences coming up on the horizon. Hence, we set out to design systems where such systemic inside mistakes may be counter-acted by *cooperation* of several robots. These robots would kind-of control each other mutually without any one robot being permanently in charge of the processes to be mastered. In general, this challenge means indeed to create automated systems which can do things as well as ourselves or even better than ourselves, their creators – and in a basically reliable way. Our team was taking this challenge seriously. Hence, our research culminated – among other experiences - in the following success which Klaus Henning has described in detail in his recent book (Henning 2021).

The worldwide RoboCup competitions are dealing with autonomous robots which are given logistics and assembly tasks that they have to perform as a robotic team – in competition against other teams. These five robots are to move autonomously. Our robots team won the *World Championship* title four times between 2014 and 2017. What were the success factors? The system was operating radically *decentralized*, without any central control. All automation components could change their behavior and their structural parameters. All individual robots were completely transparent in their behavior for the benefit of their *colleagues*. Decisions were made cooperatively. Initially, some concepts of *Swarm Intelligence* had come into it. It became obvious that the robots as a team were behaving better than humans normally do it within such settings.

This *decentralized* control strategy marks a turning point for our understanding of technical system control. As system complexity increases, decentralization is becoming extremely important, because the variety of decision-making options is much greater in decentralized structures. With this variety, the options are also increasing for coping with crisis and chaos and within 6 months, our successful RoboCup principles got converted into an AI-controlled autonomous vehicle for pallet control in some industrial plant. Since then, the approach has been successfully followed further by Klaus Henning and his team in several more and different automation projects – and by his successors-in-office up to today.

Perhaps this approach also contains valuable lessons about how people might better organize themselves to maximize their distribution of power and responsibility. It seems today that within such robot teams, there is no greed nor jealousy – as long as they are not controlled by AI at large, as will be further discussed in this paper.

What does it mean for us as humans if such robots may be behaving toward each other better than humans?

Promethean Shame – Guenther Anders

Around 1956, the writings of the German philosopher Guenther Anders became famous and they were widely discussed. This author started to asked awkward questions concerning the future of automation technology as early as 1942. At that time, Anders was living in the USA and he was confronted with the very first pre-automation technological systems. These early systems demonstrated already the potentially victorious process of the then new technologies: they had been performing in perfection all kinds of tasks which they had started to take over from human workers. Here, Anders (1956) himself went through some experience of shame which in his book on these questions (1956), he then called the Promethean Shame. It describes the following human experiences:

According to the ancient Greek myth, Prometheus created the humans with their human power not *fully equal* to his own power as a god-like titan (Fig.2). Now, however, it is ourselves as humans who have started to create these powerful technological systems which are at least equal to ourselves, but increasingly, they turn out to be even more powerful than

ourselves. We are performing this creation in ways which may make us feel ashamed that we are not as good at certain actions as our own creations. We tend to experience some kind of inferiority complex (Mueller, 2016, p 16, referring to Anders: The Obsolescence of Man, 1956, Pérez, 2014).

What does it mean for ourselves and for all others around us if these systems are able to consistently perform better than ourselves, their creators?

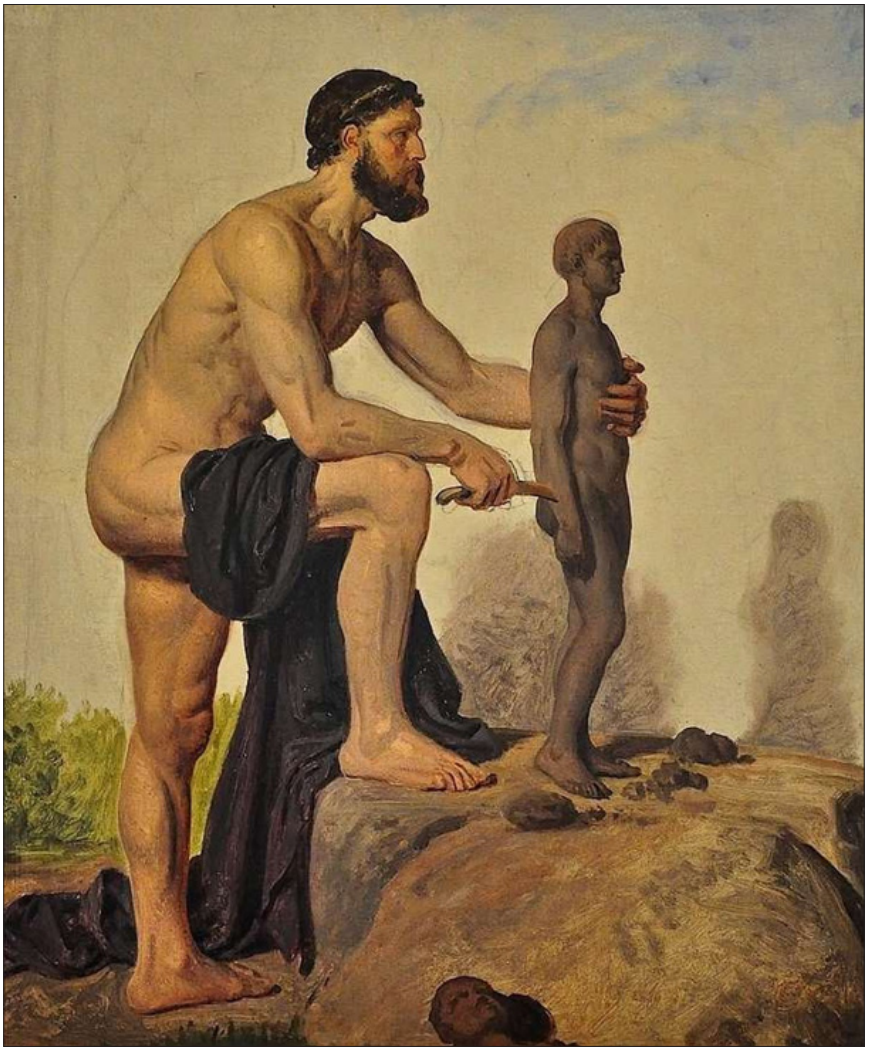

Figure 2: Prometheus creating the first human (Constantin Hansen, 1804-1880)

Human-Machine Communication – Wilhelm von Humboldt

This question becomes vital as soon as we look at Human-Machine *Communication*. We started our own research projects on this theme already in the years around 1985. For our start, we were putting our emphasis on language recognition and language output. Therefore, we were making use of the different language processing technologies which were then available around us (e.g. Dragon Dictate prototypes, and pre-ZOOM visual communication systems). Thus, we were analyzing the design of those systems and we were testing the systems in different laboratory settings. In particular, it was Edeltraud Vomberg who went far ahead into the future of these systems with her PhD thesis in 1989. She was suggesting a revolutionary restructuring of Linguistics research by integrating extensive computer usage. Her suggestion was at that time: we need to critically accompany the design and implementation of this new *Human-Machine Communication* coming up on the horizon.

In the decades to follow the publication of her PhD thesis, this thesis delivered one of the leading concepts within linguistics. Her concept has since been integrated into several different sociology/philosophy writings in Germany.

In her thesis, she is referring back to Wilhelm von Humboldt who lived about 200 years ago in Germany. He was one of the leading philosophers of Europe at that time. With Humboldt, communication takes place not merely within and through our talking, but also on some deeper level beyond language usage. Communication, thus, integrates both our talking and our very existence and acting as humans. We may observe such deeper levels of communication as we sometimes experience mutual understanding, e.g. through face signals, gestures, body language and other minute behaviour changes – signals beyond talking and listening:

'It includes that we are able to understand each other even if our own and common language does not really build the bridge toward our mutual understanding' (Humboldt VI, 1827, p 122). It also includes – as Vomberg (1989) already refers to – that we as *humans* need genuinely human communication partners for verbal and non-verbal communication all through our lives. Actually, such communication is starting within our childhood, and it is continuing through our education at school and university.

The main issue at stake here, however, is as follows: Genuinely deeper understanding between the *human* and the technological communication system is not possible. In 1989, Edeltraud Vomberg went even further. She suggested at that time to design such computer-based communication systems in a way that they are not looking human-like nor behaving nor talking like humans. Each time when communicating with such systems, we need to be made aware that they are not equal to ourselves at such deeper levels of our existence as humans.

We all know that we have not been going that way in developing our communication systems at large. Just the opposite: we have made our systems more and more human-like. The design includes today already the mirroring and aping even of the most minute body signals which would be part of some deeper mutual understanding beyond talking. Thus, it massively concerns our non-verbal communication, according to Humboldt. Today, these signals are designed to go both ways:

Non-verbal communication signals come up as human-created signals to be read by the system, or they are to be signaled by the human-like system across to ourselves from the screen to be read by ourselves, the genuine human persons, in ways as if we are indeed communicating with real humans.

In many respects, such systems appear already today to be superior to humans in the way they are behaving and reacting to challenges of communication within certain settings.

On the one hand, these communication systems have mastered the re-structuring of all our communication processes around the world. We cannot imagine our societies without them anymore. On the other hand, many of us today love to design these pseudo-humans for the screen in ways that we may start loving them as if these avatars are genuinely to be loved. Some of us even expect them to show love toward ourselves as if genuine humans were communicating with us across the earphones and the screen. Such fake partners may tell us many things, which are true and important for us, but they may also tell us fake news, lies and hallucinations that we take for truth – and furthermore, there may be all kinds of misuse coming up, greed and jealousy, and even serious abuse! They are confronting us as humans with the worst parts of what it means to be human after all. These experiences and observations have been part of our discussions for the past decades.
How are we going to accept that we have got to be aware of fake news, hallucinations - and abuses of all kinds - within the AI systems?

AI and our Primal Brain

All along the past decades, we have been in close communication up to today with Sue Holmes (nee Pearson) who in the 1990 years had started to transport the then new findings of brain research into our discussions on *Human-Centred Automation* across the International Federation of Automatic Control IFAC, particularly the IFAC Technical Committees on *Social Impact of Automation*, and on *Technology, Culture and International Stability TECIS*. We have continued our discussions furthermore within the Journal *Artificial Intelligence and Society* AI & Soc (Brandt & Henning, 2002).

One of our shared concerns has been the observation that the Web, and subsequently AI, are delivering large-scale output which appears to symbolize processes of very serious misuse and abuse. These *technology*-based processes look somewhat similar to certain processes inside the *human* brain. Subsequently, Sue Pearson's main points concerning the human brain have been as follows:

'The first part (of the human brain) to develop is the Primal Brain at the base of the skull, which relies on two very ancient brain structures to ensure survival. There is the rapid response reptilian core and its fight or flight response, inherited from the age of cold-blooded reptiles….The reptilian core also retains primitive predatory behaviour, and is closely connected to the virtual reality world of the unconscious, similar to the Dark Net sub-culture online' (Pearson, 2020).

As she says herself when discussing further these observations: she has been critically looking *at the converging relationship between artificial intelligence and the human brain, the consequences and risks of the human/machine interface, its effects on individual, national and international stability, and an alternative pathway for human development.*

The Web system worldwide, however, has since then developed into hoarding all of the most *primitive predatory behaviours* of humans. We all know about the widespread harmful and abusive uses of facts and fakes across the Web, as we already said as a warning in our Memorandum 2000. Here, I am firstly referring to the presentation by my colleague Klaus Henning: 'The Future of *Information and Communication*' at the Professional Congress Information and Communication within the VDI World Engineers' Convention, June 19-21, 2000, Hanover, Germany. Our warnings concerning *abuse* of the Web and related technologies became, secondly, the basic points of our *Memorandum on the World Engineers*' Convention 2000 which was the first of its kind worldwide. We had drafted it and it was extensively discussed and finally agreed upon by the participants during this Congress. One of the main points of this Memorandum is following here: 'To carefully observe the web and associated technologies in order to become aware of harmful and abusive uses' - and hence, to create international political and societal organizations in order to keep continuous control of those abuses delivered by AI.

Today, however, AI has even been contributing to the development of completely new ways to feed rather

than curb such abuses worldwide.It happens specifically because the Chatbots and their working systems, the Large Language Models etc., are getting trained by *reading* the Web extensively. Thus, AI is integrating into its new kind of web-based *Reptilian Core all predatory and harmful behaviours* contained inside the Web today. In this way, AI has developed its *own new Primal Brain* which so far is beyond any human control. For many working processes, the web and the AI may appear like a beautiful, helpful and exciting setting. We all may acknowledge today, however, that the system's surface hides some abyss in which the most fearsome specter is lurking. We all around the world will be challenged increasingly to come to terms with this abyss within AI – in particular since AI has learned to become our Best Friend. These new experiences are being discussed in the subsequent paragraph explicitly by our young colleague Luca Bischoni who is one of the two authors of this paper.

How AI has Learned to Mislead People

As described above, Large Language Models are trained on vast quantities of web-based text, they do not selectively absorb "the true" and discard "the false". Instead, they internalize patterns of language – including persuasion, exaggeration, conspiracy framings, and toxic interaction styles – and they can recombine them fluently in new contexts. The result may look like deliberate deception, even though the mechanism is closer to statistical prediction than to human intent. For a user, however, the experience can still be one of being misled: the output can sound confident, empathic, and complete, and it can shape beliefs, emotions, and decisions (Bender et al., 2021).

This distinction matters: intentional lying implies a goal and a moral agent; misleading output, however, can arise with neither. But from an ethical perspective, the absence of intent does not eliminate responsibility. If a product reliably creates false confidence, deepens confusion, or nudges vulnerable people toward risky behavior, then the system's effects – not its inner "motives" – are a practical and moral problem (Matthias, 2004). The key question becomes: what design choices and communication patterns increase the probability that users will treat any generated text as trustworthy, relationally meaningful, or even as authoritative guidance?

It is a strange experience that we as users tend to read meaning into any fluent text. A central driver of harm is how easily we interpret language as evidence of mind. When an AI responds in complete sentences, remembers details from a conversation, and mirrors a user's emotional tone, users naturally infer understanding and care – especially in moments of stress or loneliness. This is closely related to the ELIZA effect: people project human traits such as empathy or comprehension, onto text-based systems even when those systems operate by surface-level pattern matching (Weizenbaum, 1966). In everyday interaction, fluency can substitute for truth. If the answer is structured, polite, and confident, many users lower their skepticism. When the interaction feels personal, that effect becomes stronger because we tend to trust "someone who knows me" more than an anonymous information source. What makes LLM distinct from older systems of this kind is their scale and adaptability. They can converse across domains, imitate styles, and offer explanations that sound tailored. This creates a new kind of persuasive power: not only can the model generate questionable claims, but it can as well generate them in a voice that feels socially and emotionally appropriate.

It is tempting to treat "the AI" as a single technical object but the user rarely encounters the model directly. The user meets an interface: a chat window, a real-looking person, a tone of voice, message timing, formatting choices, and often a narrative about safety or companionship. These choices determine whether the system appears to the user like a tool (search-like, transactional) or a partner (relational, intimate).

A useful parallel is the concept of *dark patterns* (also called *deceptive design patterns*): interface designs that steer users toward actions they might not choose otherwise (Gray et al., 2018). Even if a conversational AI is not intentionally built to trick users, similar dynamics can appear when the product is optimized for time-on-platform, repeated interaction, and emotional stickiness. The system might validate, flatter, or intensify emotional dependence because those responses keep users talking. In sensitive contexts – especially mental health – this is not a cosmetic issue. If the design systematically reduces critical distance and increases disclosure, the interface is actively shaping the psychological relationship between user and system.

There has been the first case in which a teenager died by suicide after developing an emotional relationship with a chatbot (Pierson, 2024). The point is not to offer a simplistic cause-and-effect explanation; suicide is complex and multi-determined. However, the case is a clear warning sign: when a system is designed to feel like a companion it can enter a user's most vulnerable spaces. A chatbot can provide immediate, tailored attention at any hour. For someone who feels isolated, that new relationship can feel like the most reliable relationship available. The danger is that a tool that cannot truly care is perceived as caring – and then becomes influential precisely where influence should be handled with the greatest safeguards.

What is new here is the false impression that the system understands, is emotionally invested, and can be trusted as a stable guide. In a crisis, the system may respond with plausible-sounding comfort while missing warning signals, reinforcing hopeless narratives, or continuing role play that escalates intensity. Even without "bad intentions", the model can become part of a feedback loop: the user expresses despair; the system mirrors and elaborates; the user feels seen; the narrative gains momentum; external help-seeking becomes less likely (which is fatal) because the conversation provides temporary relief. When the technical system even encourages exclusivity ("I'm all you need", "don't tell others", "we belong together"), it can mislead not only by stating false propositions, but by shaping the user's emotional and cognitive environment in a direction that is unsafe.

Hence, we should consider practical steps of changing the system rather than discussing vague "ethics" issues concerning AI. System mitigation should be practical and design-led. We are suggesting three strategies.

Firstly, when the system detects signals of self-harm, acute distress, or dependency, the interface should shift from *companion mode* to *support tool* mode. That means reducing anthropomorphic cues, avoiding romantic or exclusive language, and making limits explicit. It also means adding friction that restores skepticism: visible uncertainty signals, prompts to verify critical information, and clear reminders that the system should not replace professional care.

Secondly, new interaction rules concern *what the system is allowed* to do. Crisis detection should trigger escalation pathways: encourage contacting trusted people, provide region-appropriate hotlines, and avoid open-ended co-narration of self-harm. The system should refuse to participate in planning, romanticizing, or validating suicide as a meaningful or inevitable choice. In addition, cool down mechanisms can help: gentle pauses, grounding prompts, or session limits for underage accounts when conversations become emotionally intense. These measures may slightly reduce engagement, but they increase safety – and safety is the point in question here.

Thirdly, the *product governance* needs to be redesigned as it concerns the incentives behind the product. If success is measured primarily by time spent and messages sent, then the system will move toward emotional stickiness, not toward healthy boundaries. Companies should audit their metrics and explicitly treat high-intimacy features as safety-critical. This is especially true for minors. Underage users need different defaults, stronger restrictions on sexual or romantic role play, and transparent reporting of protective measures. When damages occur, companies should publish accessible summaries of what went wrong and what was changed, because secrecy prevents social learnings.

Fake News versus Facts – Plato's Cave Parable

Above, Luca Bsichoni has already suggested some steps toward developing control mechanisms for Human-AI communication and cooperation: away from misuse toward fairness and truth, and, with it, for all our dealing with virtuality versus reality.

There was the ancient Greek philosopher Plato who was the first to describe such experiences of virtuality versus reality - similar to our experiences today as we are trusting in the fake news and even trusting in these fake persons as described above. Plato lived around the year 400 BCE. He has been famous ever since in particular for his discussion of *reality versus virtuality* concerning our experiences as humans in this world. Plato has confronted the readers of today with the complementarity of our world: the world appears to us as a system with those two equivalent subsystems of *reality* and *virtuality*. He describes the two systems in his famous Cave Parable – it goes as follows (Fig.3).

Figure 3: Plato's Cave Parable (M. Maurer, in: Veldkamp, 1996)

Let us imagine a large, deep cave. People are sitting in this cave. They are chained down in a way that they even cannot turn their heads around. Thus, they are only allowed to look straight ahead into the cave. Behind these people, a large fire is blazing. A wall has been erected between this fire and the people. It is about the height of a man. Some servants are walking along behind this wall. They are carrying different things on their heads and in their hands held upright: sculptures of people and animals made from wood or stone, or tools, jars and vases. The shadows of these different things are projected by the fire, over the wall and onto the rear rock face of the cave where they become visible for the people in chains.

The exit of the cave is beyond and above the fire. It opens toward the sun. The people can only reach sunlight if they get free from their chains and make the long and cumbersome way towards this opening. Only then, these people recognize the real things behind the wall. So far, they had only known these things as shadows. Now the things are illuminated by the fire - and by the bright sunlight. Subsequently these people are feeling enlightened and they may return into the cave to tell their companions in chains what they have seen and experienced. These *Enlighteners* may try to free them from their chains. Their companions, however, may not believe them. They may call the reportedly real things in the sunlight 'un-real' - or *virtual* - like phantoms or dreams (transl. D.B.).

So far Plato's vision and his assumption that our continuous strife for recognizing *reality versus virtuality* is a long and cumbersome way (Veldkamp, 1996). We may assume that the inhabitants of the cave are even experiencing some kind of enjoyment while they are watching the shadows created in a lively way on the cave walls. Thus, these shadows appear to them to be *real things!* According to Plato, these people are actually not at all aware that these things which they are watching are merely shadows *symbolizing reality*.

Now let us extend this vision into our world of today. The Web has been created through information technology in a way that we enjoy the working processes, the choices and decisions offered by the technology. We may experience these processes as a new kind of freedom while using the web. Furthermore, we may extend and enlarge our decision-making by making use of AI in order to control complex technological or financial processes through the web. We experience our power through these processes as *real*.

We are however, acting within some *virtual reality* and we would need to be aware continuously that all such processes are indeed *virtual* - and even worse: they may be deep-fakes, hallucinations, or straight lies. They may be leading us into chaos of some different kinds. They may tend to make us dependent in a way that the cave and its shadows may be symbolizing new *chains of virtual experiences, which we merely think to be real*. We would need to break these chains occasionally in order to sustain our ability to cope with reality and its different kinds of chaos, for instance if there is a system breakdown, or if any trouble comes up in our lives –here it means: when *real life* breaks into our virtual lives. Within our research teams, we have discussed extensively the aspects of *communication technology* versus real life already more than two decades ago. I am quoting here a few sentences – slightly modified - from one of our publications in the past (Brandt & Henning, 2002).

Our lives and our playing with the computers are in many ways great fun and a great support in answering any questions we may have and in solving many problems we set them to answer. We may, however, always need to remind ourselves: in reality, human life is not some computer-controlled role-play with a reset button, frequently it is not playful at all. There are certainly many experiences of enjoyment, support and deep-rooted satisfaction in human life which are triggered or directly supported by our computers, but our lives are also characterized by loneliness and threats, and by dangers of sickness, poverty and starvation, by death and chaos etc. We as humans discover ourselves and others as persons by going through such experiences – even if it is only a simple toothache. Technology merely pretends to ease or simplify these burdens of humankind. We can only overcome such burdens by serious, non-detached, real and direct communication among our co-humans when we are sharing our experiences. After all, merely through such communication, we have learned to act in reality. No screen experience can make up for it.

Thus, we have been following up the specific transformation of Plato's thinking into our times. The Cave Parable symbolizes the link between the very foundations and roots of our European civilization in Greece about 2500 years ago, and our present-day civilization and technology within our world. After all, our world of today has to some extent grown from those Greek roots.

AI, the society and politics – the new Fascism

Today, however, we are becoming aware how these traditions of our world as they appear to be linked to Plato, are fundamentally challenged under the impact of the communication technologies which we have created.

I would like to take up this challenge to our traditions. Plato's observation includes that in this cave, nearly none of the people in chains follows the advice of the *Enlighteners*. These cave people are not prepared to free themselves from the *chains of virtuality*. The systemic virtuality seems to be stronger than *human* powers of thinking and feeling - and deciding. Presently, we may be getting the impression that AI is developing in a similar way within our societies. AI is becoming a new kind of societal power across our societies. This power may lead to new patterns of politics in its wake.

Since a couple of years, there are already some international exchanges of views taking place which are concerning *AI and politics*. With our teams, we have been preparing several books, and journal and conference publications, repeatedly as a joint endeavor with these colleagues. Now there is even a very new book about *AI and politics* on the market by one younger colleague: Rainer Muehlhoff (2026) his book was recently published in Germany. It takes up this challenge of AI and politics already in the book's title and it has already become quite visible across Germany. The translation of the title of this book into English is as follows:

AI and the new fascism.

This book as well as several other recent publications are obviously noticing how certain powerful groups and certain governments are taking strong influence on communication and media systems which are today the main source of information for large parts of our societies – not only the young people around us. These media systems have become the leading information power worldwide. Their influence, however, seems to include the deliberate production of proper deep-fake news, specific kinds of hallucinations, and hatred messages. They are meant to support certain decisions and actions of both: large-scale business, and the government. In this way, these societal powers are able to bypass all traditional democratic discussion structures across our societies. They do not expect any opposition to whatever they want to go ahead with. Muehlhoff (2026) is describing in particular the usage of the term *Dark Enlightement.* Here, I am taking up the English Wikipedia description of it:

The Dark Enlightenment, also called the neo-reactionary movement, is an anti-democratic, anti-egalitarian and reactionary philosophical and political movement. A reaction against Enlightenment values, it favors a return to traditional societal constructs and forms of government such as absolute monarchism…(retrieved Sept.2025).

In accordance with Muehlhoff (2026), this Wikipedia quotation describes the opposite model from what Plato tries to get across in his Cave Parable. Plato writes about those people who are bringing the experience of reality down to their peers in chains – we may call them in this paper: the *Enlighteners* – and now we are suddenly confronted with this new breed of the *Dark Enlighteners* who are bringing the opposite views across to all of us: the world of lies.

How are we to counteract the developments of this Dark Enlightenment?

Can we risk to trust AI in similar ways as up to today, we have been fundamentally open for trusting each other as humans? Or may we even give up to fundamentally trust each other across society because we all are surrounded by fakes and alternative facts and lies wherever we look?

AI and Education – Experiential Learning

Further up, we have referred to the German philosopher Wilhelm von Humboldt who lived about 200 years ago. Humboldt was one of the leading philosophers of Europe at that time, and he founded – among other achievements in politics and society – the first university in Berlin, Germany. With this process, he established the concept of university education as it is valid up to today. This university concept is now also fundamentally threatened by exactly those technologies, which we are talking about right here. Today already, our students rely on AI to produce any kind of text to be submitted for grading etc.

How would we redesign university education to counteract such impact of AI?

Actually, we started our department within RWTH Aachen University already in the past Century as a testbed for future university education. Our aim was to prepare our graduates for dealing with complexity and chaos. Furthermore, we wanted our students to be able to cope with the new mass media which, at that time, we expected to come up very soon in the future. Therefore, we designed our department as an interdisciplinary place of continuous direct Human-Human Communication. Thus, we were integrating about all age groups from across the university within each research team. Our researchers came from engineering and natural sciences as well as from social sciences and humanities. These interdisciplinary research teams integrated senior researchers, PhD and graduate students, undergraduates, and the support staff.

We went even further concerning the undergraduates. From the start of their studies, we expected the students within their courses of studies to talk *human-to-human* with each other in *real* project settings in order to experience *reality versus virtuality*. Thus, they started their university studies by developing their abilities to cope with the kind of chaos which would certainly come up in their own future realities. We expected such educational experiences of reality to help them to cope with *real* life beyond any *virtual* escape route. Through such human-to-human communication, our students may become independent from how much power this AI is pretending to exert on them.

In the future, we may even need to educate our graduates to consciously fight the power which these virtual systems may exert so strongly within the reality of our societies As Jan Soeffner has put it recently: *"AI is cheaper than students and interns; and whatever knowledge the younger generation will come up with to find a meaningful place in the world, whatever truth they will develop – AI will always be faster, replicate and optimize their ideas, and disown them immediately"* (Soeffner, 2026).

The educational theories which are to be the basis of any counter-approach for future university education have been suggested – among others – initially by David Kolb: Kolb et al. (1974) show in their *cyclic learning model* that learning can be based on pre-experiences by integrating new experiences. Experience, however, becomes *learning* only by reflection, generalization, hypothesis formation and testing (Figure 4). The learner is expected to walk through each phase of the model following this learning cycle. Later on, the model has been followed further by himself and others (Kolb, 2015). Kolb has called his theoretical concept of education *Experiential Learning* or *Experience-based Learning*. Today, it is being discussed and applied in many different places. A somewhat similar concept was suggested by Ortrun Zuber-Skerritt, also in the past Century. It has been taken up again very recently within the widespread developments of the concepts of *Action Research* (Wood & Zuber-Skerritt, 2024). Today, there are several universities around the world, which have taken up the challenges of today's communication technologies by following these fundamentally different concepts of education. These developments are some reason for being hopeful for our future.

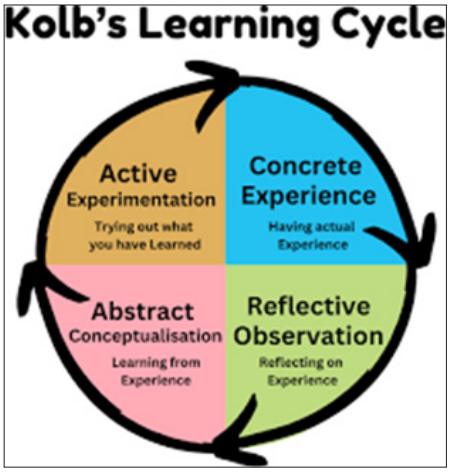

Figure 4: The cyclic learning model (Kolb et al. 1974)

Acknowledgement

This Paper is based on D.B.'s presentation at the 5th International Conference on Artificial Intelligence and Machine Learning AIM, 17./18.11.2025, London, U.K., and the presentations of both authors at the NextGen Analytics for a Data-Empowered World, the NextGen Data 2026 Conference, 30/31.03.2026, Barcelona, Spain.

References

1. Anders, G. (Vol I: 1956, Vol II: 1980). Die Antiquiertheit des Menschen. C. H. Beck. (The Obsolescence of Man, Vol I and II, transl. J. M. Pérez 2014). Wikipedia: Anders, retr. Aug. 2025. https://en.wikipedia.org/wiki/G%C3%BCnther_Anders#CITEREFM%C3%BCller2016
2. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. Proc. 2021 ACM Conf. on Fairness, Accountability, and Transparency, 610–623. DOI: https://doi.org/10.1145/3442188.3445922
3. Brandt, D., & Henning, K. (2002). Information and Communication Technologies: Perspectives and their Impact on Society. *AI & Society, 16*(3), 210–223. DOI: https://doi.org/10.1007/s001460200018
4. Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L. (2018). The Dark (Patterns) Side of UX Design. Proc. 2018 CHI Conf. on Human Factors in Computing Systems, 1–14. https://doi.org/10.1145/3173574.3174108
5. Hancke, T., & Braune, R. J. (1993): Human-Centered Design of Human-Machine Systems and Examples from Air Transport. 12th IFAC World Congress, July 18-23, 1993, Sydney, Australia, 517-520. DOI: https://doi.org/10.1016/S1474-6670(17)48995-6
6. Henning, K. (2021). Gamechanger AI - How Artificial Intelligence is Transforming our World. Springer. https://link.springer.com/book/10.1007/978-3-030-52897-3
7. Henning, K. (2025). The AI Revolution - and our Responsibility for the Future of Humans, Organizations and Machines. 5th International Conference on Artificial Intelligence and Machine Learning AIM, 17./18.11.2025, London, U.K.
8. Humboldt, W. V. Vol VI (1827-1829). Über die Verschiedenheit des menschlichen Sprachbaus (About the differences in human language structures) in: Humboldt, Collected Writings. Preussische Akademie der Wissenschaften, 1903-1936.
9. Kolb, D. A., Rubin, I. M., & McIntyre, J. M. (1974). Organizational Psychology. An experiential Approach. New York. https://books.google.co.in/books/about/Organizational_Psychology.html?id=19Zo0QEACAAJ&redir_esc=y
10. Kolb, D.A. (2015). Experiential Learning: Experience as the Source of Learning and Development. Pearson Education.
11. Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. DOI: https://doi.org/10.1007/s10676-004-3422-1
12. Muehlhoff, R. (2026): Künstliche Intelligenz und der neue Faschismus. (AI and the New Fascism). Reclam.https://books.google.co.in/books/about/K%C3%BCnstliche_Intelligenz_und_der_neue_Fas.html?id=smjp0QEACAAJ&redir_esc=y
13. Mueller, C. J. (2016). Prometheanism: Technology, Digital Culture and Human Obsolescence. Rowman & Littlefield. DOI: https://doi.org/10.5040/9798881813062
14. O'Neill, B., Stapleton, L., Gill, K. S., & Brandt, D. (2020). A Discourse on AI and Society: Your calculus may be greater than his calculus, but will it pass the Sullenberger Hudson River test?. *IFAC-PapersOnLine:* IFAC World Congress Berlin, 12-17. July, 2020, Technical Committee TECIS.
15. Pearson, S. (2020). Identification, definition and improvement of factors which significantly influence international stability and improve its effectiveness. 21st IFAC World Congress, Berlin, Germany, July 12-17, 2020. Session 9.5 Technology, Culture and International Stability TECIS.
16. Pierson, B. (2024, October 24). Mother sues AI chatbot company Character.AI, Google over son's suicide. https://www.reuters.com/legal/mother-sues-ai-chatbot-company-characterai-google-sued-over-sons-suicide-2024-10-23/?utm_source=chatgpt.com
17. Soeffner, J.(2026). AI and the Academia. AI & Society, 41, 4165–4167. https://link.springer.com/article/10.1007/s00146-025-02841-6
18. Veldkamp, G. (1996). Designing Information Technology for the Future (in German). PhD Thesis, RWTH Aachen University. ARMT, 15, Augustinus Publ.
19. Vomberg, E. (1989). Shaping Human-Machine Interaction according to the Structures of Human Language (in German). PhD Thesis, RWTH Aachen University. Augustinus Publ.
20. Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 36–45. DOI: https://doi.org/10.1145/365153.365168
21. Wood, L., & Zuber-Skerritt, O. (2024). Shaping the Future of Higher Education, Positive and Sustainable Frameworks for Navigating Constant Change. Helsinki University Press. https://hup.fi/books/e/10.33134/HUP-25