Götz G. Wehberg

Human AI

What it takes beyond trust to keep control over AI and run it beneficially (Part 1)

DSI | DIGITAL SUPPLY INSTITUTE

Human AI
What it takes beyond trust to keep control over AI and run it beneficially (Part 1)
Götz G. Wehberg

First edition published 2023
by DSI – Digital Supply Institute
Amsterdam / Shiphol, The Netherlands

DSI DIGITAL SUPPLY INSTITUTE

ISBN: 978-3-00-077618-2

Foreseen to be published in 2024:

Human AI - What it takes beyond trust to manage AI beneficially (Part 2)

A #DigitalMadeInEurope initiative

Content

Preface

Trust is necessary to collaborate between humans as much as with machines. And collaboration is what allows us to develop and create civilization, wealth and progress. Nothing works without trust. This is particularly worthwhile to mention, given that there is a current lack of trust when it comes to artificial intelligence (AI) and the internet as the facilitating platform for AI. Many politicians, scientists, tech experts and managers are therefore calling for AI to be trusted. But is this good enough to secure AI is beneficial to humans? I am afraid it`s not.

We cannot simply trust AI unreservedly and hope that AI will help us humans. The call to trust AI can be misunderstood as a plea to please not interfere and that few global tech companies are already doing that. This essay, therefore, tries to determine the guiding principles to make sure humans benefit from AI. While trust is a necessary condition, these guiding principles are supposed to be the sufficient condition for its beneficial use. And these principles are facilitating the build-up of trust, just because they ensure AI is both trustworthy and beneficial.

Now, how to make best use of AI for the benefit of humans. Three challenges lie ahead of us:

1. Can we regulate digital technology and AI via good governance to make sure humans keep control?
2. Can we run AI, operationally, in a way that it really delivers expected benefits to humans?
3. Can we manage tech organizations in a way that indeed appreciates humans and fosters beneficial digital innovation as actual outcome?

The good news in brief is, yes, we can. But how? And yes, there will be more questions beyond these three. But let`s start to make a double click on these three perspectives of Human AI, and then we can add as we go.

How can we regulate digital technology via good governance for the benefit of humans?
When it comes to digital technology and AI, we must recognize that it is increasingly possible to place our knowledge and technological know-how into machines. This is a fundamental change because so far millions of IT managers have lived on this planet. They have probably spent on the order of dozens of million person-years learning and teaching, so that our IT can be maintained and run. In the future, super-intelligent systems - in the sense how for example the Berkley professor Stuart Russel describes them - will be able to run IT systems for us and even manage our lives. These kinds of systems are capable to comprehensively store and apply know-how, which is vital to maintain and develop our civilization. Therefore, we need to ask ourselves how to make best use of such super-intelligent systems and AI in the future.

AI is not supposed to hold authority over humans. Neither so, are single tech companies that own the globally leveraged AI platforms today. There is a chance to develop proper AI regulation, i.e., making sure AI benefits humans. Let`s call this "Human AI". The current AI Act of the European Union is a good first step. The time is now as super-intelligent systems, which will provide Everything-as-a-Service with the potential to either dominate or benefit humans, significantly, are developing fast.

This essay outlines the gaps or white spaces with respect to Human AI regulation, the guiding principles and next steps to get there. Over and beyond good governance, it is in the hand of humans to run AI systems in a beneficial way. This leads to the second question, which is not less complex than the first one.

How can we run AI systems so that they deliver expected benefits to humans?
To run digital technology and AI systems in a beneficial way, one must understand what`s new. Therefore, the essay summarizes the differences from a technological as well as functional standpoint. LLM-based operations (LLMOps) is leaping technology, away from so-called rule-based coding towards imitating best-practice behavior, comprehensively. To run AI in this way, brings up practical questions like who is supposed to define best-practice, how to stay agnostic from single foundation-model vendors and how to maintain transparency, ertc. Guiding principles with respect to AI operations add up to the principles of good AI governance, accordingly.

While the answers to the first two questions are being outlined in this Part 1 of the essay, Part 2 deals with the third question, which is not less important. In Part 2, we focus on the management question:

How to manage digital technology with means of a human tech organization that fosters beneficial innovation? Digital technology and AI are creating so many expectations, some realistic, some futuristic. So, how can tech organizations deal with all this in a human way and how should CIOs set up to drive business innovation?

The discussion of the answers to the three questions clearly shows that it cannot only be the task of politicians to ensure the correct use of AI, even if regulation must play a decisive role. Rather, it is also the task of AI users and providers as

well as CIOs, IT experts, auditors and other stakeholders. I am convinced that we are all aware of this. Human AI is a societal task that we must all face up to. It is a must to be proactive and help shape the future instead of being reactive and too late. I would like to thank all those who have accompanied me and those who are interested. I look forward to exchanging ideas, arguing on the matter and continuing to learn together!

Götz G. Wehberg

1 How humans maintain authority over themselves

1.1 Tech companies taking the wrong turn

Digital technology as well as AI are imposing significant challenges. Rather sooner than later, digital technology is leaping towards super-intelligent systems, which will be able to provide solutions as a service for everything humans ever imagine. Such super-intelligent systems are supposed to develop solutions for the biggest problems on earth such as global warming, over-population, scarcity of resources and health equity amongst others. Humans can, therefore, very much benefit from super-intelligent systems, however, such systems can also destroy humans as a species in a worst-case scenario.

Let`s illustrate the future a bit more. Super-intelligent systems will enable personal digital assistants that might make our life easier. Personal digital assistants that we can trust might use our bank account, reply to mails and answer phone calls on behalf of us. But do we want personal digital assistants to vote for us in elections, share information about our mental health status, choose medicine or make the call for therapies. Probably not, you might think. You want control about it. But how do we make sure we keep control as humans and maintain authority, while making best use of this new technology. What we can`t do is just waiting until the technology has developed and either scenario becomes reality. Because then it might be too late.

When it comes to AI, I have mentioned already that some leading tech companies have taken the wrong turn in the past. These companies always assumed the more intelligence a system is the better it can meet their purpose. AI systems optimizing themselves, however, do not necessarily consider the benefit for humans. Many of the AI systems have been instrumentalizing humans with means of choice-based preferences learning mechanisms for their own purposes, e.g., maximizing ad revenues. To get as much attention as possible, these AI systems have developed negative narratives, fueling fears and hate. They have reinforced the perception of what people wanted to believe anyway, which has led to many living in their own bubble. For not just a few people, this has reinforced an narcissistic, anti-democratic and even racist attitude, which is the opposite of a fact-based, open-minded coexistence in a diverse world. What we must learn from this is that whenever AI-based services are biased by their own purposes or associated with separate financial or commercial ones, they can become independent from us, establish their own authority and create a conflict of interest. This is basically true for AI-based services in all sectors and all realms.

The issue comes from the fact that super-intelligent systems are capable to outperform humans. A super-intelligent system would then be subject to Goethe´s “Zauberlehrling”. The proverbial phrase "I can't get rid of the ghosts I called" vividly describes the dilemmas of an AI science that will be held responsible for the consequences of its research. There is common agreement one should avoid putting a proprietary purpose into the system. Proprietary purposes can range from “making our life happy” to “choose the pizza we want”. Fair enough, you may hold against this that it can be doubted whether humans themselves are able to determine what makes their life happy. For example, if all work is done by AI, are we happy without any work or do we need some to give our life a purpose. And is it necessarily all the same for every human? Indeed, humans do not know. So, how would machines know.

Moreover, even lower preferences like choosing the right pizza must not be pursued by machines at all costs. If the pizza restaurant is closed, do we want the AI bot to drive 5,000 miles to the next restaurant or not. If not, where is the limit and how long are we willing to wait. Even such simple preferences and the constraints associated are quickly becoming complex and are often not easy to determine by a machine without asking in that specific situation.

That said, I would like to refer to Paul Daugherty and James Wilson stating: “Business that understand how to harness AI can surge ahead. Those that neglect it will fall behind.” I basically agree with Paul and James. You just must read “Human AI” not just “AI”. Otherwise, you run at risk to fall into the same trap as leading tech companies have done in the past. Now that the first wave of enthusiasm about AI is snapping a bit and many people are beginning to understand its possibilities, we need to take a sober look at AI and what needs to be taken to heart for this technology to really benefit us humans. So where do we stand today?

While named tech companies are trying to fix this, their current approach is again not necessarily going into the right direction. Let`s call these new efforts the “Behavioral Approach”. The Behavioral Approach suggests that future AI systems can be “provably beneficial” by applying three principles: By aiming only at the realization of human preferences, starting with an initial uncertainty what those preferences are and predicting human preferences via observed behavior, which can be expressed in actual human choices. As the Behavioral Approach suggests, key industry players like Apple, Baidu, Meta, Microsoft, Tencent amongst others should partner on AI topics and self-obligate towards these three principles. Representatives of the Behavioral Approach admit, the approach should be considered as an ambition rather than a guarantee.

The Behavioral Approach, though, doesn`t say so much how AI is governed effectively. Behavioral sciences build on observed behavior of humans. These sciences, however, are far away from being able to predict real human preferences, whether supported by a machine called “super-intelligent system” or not. Of course, it makes a difference whether you try to predict the choice of a salami on somebody`s pizza opposed to real human preferences or even a political will. While understanding the methods of choice-based preference learning and statistics associated (like conjoint analysis etc.), it is important to recognize their limits. Sciences that try to predict human behavior were moving away from Stimulus-Response models that work for Pawlow`s dog towards Stimulus-Organism- Response (SOR) models, which are by far more complex. Yet even SOR models do not grant predictability in many instances. For example, a human can be environmentally conscious, but doesn`t act environmentally friendly in some cases. AI systems, which refer to the Behavioral Approach, though, would conclude the human doesn’t care for sustainability or doesn’t have a green attitude. Of course, behavioral sciences develop, but how long will it take. Even if behavioral sciences deliver on time, which I doubt, the proposed principles of the Behavioral Approach can be operated like a proprietary purpose themselves, thus repeating the earlier mistake.

Therefore, AI is not supposed to hold any authority over humans, including predictions of human preferences. The speed of AI development over the past years was significant, e.g., the achievements of OpenAI. It must be doubted that behavioral sciences can offer solutions before super-intelligent system become a reality. The core challenge is that no system can define true human purposes, which leads to a misalignment with our values. Self-obligation is not good enough as it hasn’t worked in the past. And competition between companies, nations and regions motivate incompliance. Take the navigation app for your car for example, which you can only use if you switch on speech recognition so that the provider can scan your conversations and your profile (observed) consumer preferences. You can say this is one`s free decision to use the navigation app or not. Practically, it ends up in discrimination of those humans who are not willing to share their data, and thus do not get access to latest technology.

Forcing people to share data that they don't want to share crosses boundaries. Spreading bad or even fake news just to attract people's attention to obtain data also crosses borders. But the Behavioral Approach needs data and access. The three principles of the Behavioral Approach, therefore, assume access to personal data of humans to facilitate the prediction of their preferences. This approach is at least close to social scoring because it classifies people based on behavior. It is also remarkable that the Behavioral Approach talks about “provable benefits” considering that Sir Karl Popper`s science theory suggests you can’t prove an empirical theory right.

The currently discussed Behavioral Approach as described above can probably work for meeting consumers` interest in line with their observed usances in a very limited scope. At its core, the approach focuses on what is being liked and expected. The real preferences or even political will of humans are much more comprehensive, deeper and complex.

1.2 Good governance for Human AI

Instead of following a Behavioral Approach I suggest the following guiding principles for a proper governance. I call this a Human AI approach. Human AI takes care digital technology is benefitting humans and governed in line with ethical standards, which protect the basic values of human life:

Before I share guiding principles for Human AI, let me mention what Human AI is not equal to, i.e., for good housekeeping I am sharing some nomenclature in brief: By Human AI, I don't necessarily mean human-like AI as Russel and Norvig (2013) understand it. Russel and Norvig differentiate between human-like AI and rational AI. The former is about the look, the mediated perception and the behavior of machines that resemble the appearance of humans and can seem very clunky if they are not human-like. In simple words, human-likeness is about the look and feel of AI. Human-like AI can form malleable habits, e.g., a driverless car picking up the local driving culture. It can be hallucinating but communicate in a very empathic way. This is not Human AI. In a different scenario, AI can make sure it is benefitting humans, but in a very clunky style of behavior and communication. Still, we talk Human AI in this case. Human AI can be human-like or not, however, these are different questions. Of course, human-alikeness can be important, e.g., when it comes to care giving robots in a rapidly ageing society (see for example Sam Freed 2022). Nothing wrong with Human AI that is also human-like, which is in fact something we expect when talking about super-intelligent systems. However, a human-like AI, which is not Human is nothing the following guiding principles are aiming at.

The first guiding principle of Human AI governance is **Independence**: A proper AI regulation must distinguish between the governing and the governed system. In other words, the borderlines between the object and subject of regulation must be clearly defined. Regulating and regulated system must be demarcated from each other to build trust, take objective decisions and prevent us from conflicts of interest.

The relevance of Independence or independent governance as a guiding principle is not limited to digital technology and AI, of course. It is known as a principle in many contexts. In politics e.g., politicians must not liaise without limitations with industry representatives. Lobbyists must provide transparency about their engagements via a lobby register and vice versa. In the context of AI, however, applying independence seems to be a bit more complex. Imagine for example, a regulatory body that authorizes the use of patient data for research purposes is supported by the same AI (operator) that markets the algorithms derived from that kind of data. This is where AI crosses the boundaries of Independence.

You say this barely happens? Look at Ashton Kutcher for example and his creditable initiative against the abuse of children. Kutcher has been lobbying for scanning mails as well as messages and has been supporting the drafting of the Child Sexual Abuse (CSA) regulation of the EU. Critics describe this proposal as “chat control”. Kutcher also has been the founder and running the tech provider for doing the scan, called Thorn. Preventing child abuse is, of course, desirable and using scanning technology can be an option. But why does it have to be the same lobbyist's Thorn software. This can clearly breach Independence.

Both AI systems and operators must refrain from liaising with AI regulators in an unguided manner. This does not mean that politicians cannot use AI for example for simulations and support of their political projects. The supporting AI, however, must be clearly independent from the AI systems which are subject to the regulation. Again, you may say that this is obvious and well known as guiding principle, but this is not what is necessarily happening today. To understand digital technology, big tech companies and AI operators are often being asked for their advice and granted a “chair at the table”. Independence is not necessarily the current focus. As AI is currently evolving fast and real expertise is short, Independence is a principle not easy to consider. But it is still necessary. In simple words, you must not ask the frogs if you want to dry out the pond.

The second guiding principle of Human AI governance is **Autonomy**. Autonomy of humans means that the final authority how to meet one`s own preferences or political will must stay with humans. That includes decisions about human preferences itself, which are not supposed to be outsourced to machines. Machines that make decisions about humans do violate their dignity. Human dignity, however, must be regarded as inviolable. For this reason, we must avoid putting a proprietary purpose into the AI system, thus stay autonomous.

Aligning between preferences and political wills of different humans is at the heart of politics and not subject to machines. Humankind and its thought leaders have given lots of thoughts over the past centuries how to ideally set up government. For Aristotle, e.g., the best form of government is ultimately the “polity”, whereby the translation into a distinguished structure hasn’t been finally specified by him. Recognized fundamentals in this context are the democracy, the inviolability of human dignity, the rule of law as well as checks and balances. The latter refers to the horizontal division of decision making between legislative, executive, and judicial branches, which is subject to the first guiding principle of Independence. Autonomy suggests in this context that decision making stays within the democratic process of elections by humans rather than outsourced to “machines”. Super-intelligent systems and AI must support these fundamentals. AI regulation must make sure, accordingly.

But of course, Autonomy is not limited to above fundamentals. Proprietary purposes that can be put into a machine range from, e.g., “making my life happy”, via “staying healthy” and “being mobile comfortably” to “choose the pizza I want”. In all instances, machines must ask or support, not conclude by themselves. As mentioned before, we must avoid putting any proprietary purpose into the AI system. Thomas Chamorro-Premuzic suggests: “Where will humans go? We should not decide on the basis of where AI can take us, for we are still in the driver`s seat, even if it often doesn`t feel like that.” I agree. We have it in our hands to become a better version of ourselves, without or now with the help of AI.

A third guiding principle is **Auditing**. AI must be audited as aforementioned principles of Human AI governance are not necessarily being met by AI providers, voluntarily. AI audits must be performed to check compliance and maintain a four-eyes oversight. For auditing, there must be a distinguished supply of independent AI experts, assurers or auditors available who can perform such AI audits. While there is no universal audit checklist yet, audit criteria currently being discussed are based on the collective experience of developers, IT assurance experts, data scientists and machine learning engineers and prevent previously known risks and errors. Gartner for example, suggests the help of a comprehensive AI Trust, Risk and Security Management (TRiSM) program. TRiSM aims at proactively ensuring that AI systems are compliant, fair, reliable and adhere to appropriate privacy policies.

Current audit efforts like TRiSM and HCAI focus rather on trustworthiness of AI, e.g., keeping private data private. They are not necessarily focusing on the benefit of AI for humans, which of course is much harder to audit. As said before, trust is necessary to collaborate, between humans as well as between humans and machines. But trust alone is not good enough. Therefore, AI audits must go beyond transparency and legal compliance in terms of TRiSM requirements. They must also provide evidence for AI systems respecting Autonomy. For doing and enabling so, Auditing must include:

- Explainability: AI systems must be inherently explainable, enabling auditors and users to comprehend the reasoning behind specific outputs.
- Traceability: AI systems must include mechanisms to trace and record the decision-making process of AI systems. This traceability aids in post hoc audits and investigations by independent auditors in line with the first guiding principle.
- Human Oversight: AI systems must integrate human oversight into the processes. This involves having humans in the loop to review and validate AI outputs in line with the second guiding principle, especially in critical or complex decision-making scenarios.
- Continuous Monitoring: AI systems must consider and implement ongoing monitoring systems to track the quality of AI models over time. A proactive approach must enable continuous quality improvement as well as the identification and mitigation of issues before they escalate.
- Education and Awareness: AI providers must foster awareness among users, developers, and auditors about the capabilities and limitations of their AI systems. Education is key to making informed decisions and conducting effective audits.

The issue with audits is that you easily go through a list of checkpoints without really understanding what the machine is doing. Take explainability as an example (see also Shneiderman 2022). Let`s assume an auditor is reviewing an AI for breast cancer diagnostics. An auditor may find explanations for certain diagnoses provided by the systems. Auditors themselves only looks at formal requirements, as they typically are not experts in the audited field, like breast cancer. In this scenario, the auditor thus ticks off the checklist. Whether the explanations are really supporting humans` decision making or not is difficult to evaluate. But what if the machine just aims at generating revenue for the hospital or maximizing the utilization of the magnetic resonance tomography system. Therefore, auditors must work together with both software engineers as well as - in this case - practitioners to really understand possible biases of the code. IT assurance standards for AI in the sense of Human AI, here specifically on Autonomy, must be developed, which is a challenge for the inter-disciplinary skill sets required.

Therefore, also audits can fail. Principles of Human AI can be undermined by AI even when audited. Remember the discrimination mechanisms described in section 1.1. We cannot exclude that AI can enforce decisions or pretend compliance with auditing criteria. The more we ask few AI systems to manage our lives, jobs, health, mobility, etc. the more accumulated power and influence they get, if we permit. For any regulation, therefore, it is not good enough to only suggest that “the Independence and Autonomy of decision making must be considered and be supported by Auditing” but it needs to be operationalized and underlined with practical mechanisms and regulation that really works. Essentially, it must balance power and influence of AI systems and their providers. To divide and conquer is the only way to practically make sure the first three guiding principles are basically being respected. This is even more relevant the more you take possibly unfair and self-optimizing behavior of a super-intelligent system and their providers into account. This is of course something we don`t necessarily assume, as we want to discuss in good faith, still not naive.

The fourth guiding principle for good AI regulation, therefore, is **Competition**. AI governance must secure the freedom of humans to choose between alternative AI systems and providers. This is ultimately a question of market structure and antitrust. Antitrust must determine how much power an AI system and its operator is supposed to accumulate. The effectiveness of other guiding principles such as Independence, Autonomy and Auditing ultimately depends on such power and control.

Enabling a balanced market structure is of course not limited to antitrust. Industry politics of different countries or association of states can support this, e.g., by facilitating industry champions. I expect that for the key AI areas like foundation models themselves as well as areas of appliance such as Mobility, Health, Communication, Entertainment and so forth we must develop a balanced supply between key economic areas like the US, China and the EU. This seems to be an underestimated challenge, probably because it is easier for politicians to impose bans than to promote innovation.

The current discussion is a chance for every state or region to develop its own regulation in terms of a Human-AI approach based on the guiding principles. Developing such regulation does not start from scratch but must consider existing regulation and current laws. Such legacy regulation must be properly adopted where necessary. It must be applied within the AI context. This includes regulation for the liability of AI systems as well as data privacy, intellectual property and cybersecurity associated.

The current AI Act as well as Digital Markets Act (DMA) of the European Union are careful first steps towards Human AI. Both Acts are far from perfect, but still better than nothing. For example, the AI Act bans social scoring. Services affected by the DMA are from Alphabet, Amazon, Apple, Bytedance, Meta and Microsoft. These include the search engine Google, the online retailer Amazon, the Windows operating system and Apple's App Store, as well as the social media app Tiktok and the video platform YouTube. The time to further develop this regulation as well as market structures is now as super-intelligent systems are developing fast.

Europe is supposed to play a key role with respect to digitalization. Basic concepts like democracy have been developed in Europe. The successful digitalization as well as consequent implementation of guiding principles that secure human`s benefit can make a difference for the future of Europe.

Of course, AI and AI players act globally. This means AI regulation must synchronize around the globe. Take the example of nuclear science, where the global state community managed to align on worldwide standards for nuclear technology. Why wouldn`t we regarding AI. Let`s be confident but alerted to make sure the right guardrails are set for humans to maintain authority over themselves. Obviously, relying on regulation alone is not enough. Organizations need to operate AI systems in the right way to ensure the benefits for people. This is what the next section is about.

2 How to run AI beneficially for humans

2.1 Complexity challenging so far standard operations procedures

We must run AI systems, functionally and technically, in a way that they deliver expected benefits to humans. For doing so, it is helpful to understand how we have run IT systems and data analytics so far, in a time before AI. Let`s take some examples to illustrate the difference, like from the Pharmaceuticals, Health Care and Automotive industry.

Shopfloor operations in the Pharma industry are typically being addressed with standard operations procedures (SOP) and quality parameters to secure Good Manufacturing Practices (GMP). Process analytics have tried to understand the stochastic part of it and increase transparency where typically judgment of the operator and thus heuristic decision making comes into play. Many consultants have spent decades to make such heuristics of the Pharma shopfloor explicit. Software companies have tried to translate such heuristics into proper coding, with mixed outcomes dependent on the complexity of operations. The current rise of individualized medicine like cell and gene therapy (CGT), however, is challenging the existing landscape of operations procedures and tools even more.

Similar in Health Care, where standard procedures are expected to secure the quality of both diagnostics and treatments, just because not every physician and nurse shares 25 years of experience or has the entire compendium of scientific research results handy, e.g., in Nephrology or Cardiology. Standard procedures have helped to secure effective care delivery at scale, i.e., within the entire team of a treatment area. As the shortage of experienced labor force is becoming more serious, quality assurance of care delivery is being challenged. Moreover, diseases like multiple sclerosis (MS) are big "drawers" in textbook medicine. There is no one simple form of treatment that covers it all. Rather, patients must be diagnosed and treated individually, where possible in an outpatient mode. Such individualized treatments and the higher complexity associated ask for an even better qualification of staff and quality control.

2.2 Rules versus self-organization

Typically, the approach to develop and define operations procedures has been rule-based in the past. For example, clinical trials explored the right medicine and scientific research offered evidence-based practices. Corresponding algorithms typically have been articulated in a kind of if-then instructions. In Health Care delivery, many scientists and doctoral students have been working to explore such law-like relationships and validate them using multivariate methods of statistics. Similarly in the Pharma operations, where supply chain managers and shopfloor operators have referred to rules of lot size optimization, sequencing, network and material balancing and so forth.

A weak point of such rule-based approach, however, is its limited potential to cope with the upcoming complexity and individualization. There is a seeing-knowing gap to recognize existing complexity as well as a knowing-doing gap to manage complexity in the right way. You can refer to Ashby`s law of “requisite variety” suggesting that only complexity “eats” complexity. This is exactly where the two types of gaps are coming from.

In Health Care for example, the technology to recognize complexity (and thus close the seeing-knowing gap) with means of better diagnostics tools, digital twins and sensors has developed rapidly. E.g., computed tomography (CT) helps to diagnose the brain, skeleton and internal organs and tracking & tracing tools help to understand the structure of the Pharma supply chain network.

For managing higher complexity (i.e., closing the knowing-doing gap), requirements have always been clearly defined in terms of self-organization, however, there were little tools and solutions, which practically helped to implement. In shop floor operations for example, first attempts refer to Wildemann’s Modular Factory, Warnecke’s Fractals, then Ptak`s DDMRP and the begin of Machine Learning (and I also tried to contribute with some earlier publications like my “Digital Supply Chains”). Proper technology to manage higher complexity in a self-organized fashion and at scale has been pre-mature for long (for more see for example Wehberg 2015 and 2021).

2.3 AI and LLM based operations (LLMOps)

This is where AI and large language models (LLM) come into play. Can LLM help securing operations quality in the future? Can it make self-organization happen, close the knowing-doing gap and thus manage the higher complexity in Health Care delivery, Pharma operations and other industries? In brief, yes it can.

As LLM-based Operations (LLMOps) seem to be the elephant in the room, let`s see what is it exactly and how does it work (for more on LLM see for example Timothy B. Lee and Sean Trott 2023). Let`s have a quick look into Autonomous Driving and how the Automotive Industry uses LLMOps. Autonomous Driving has been evolving from a rule based to LLM based system too, like Tesla`s Full Self Driving (FSD) technology (for FSD see for example Walter Isaacson 2023). FSD 12, was based on a new concept that Tesla believes will transform autonomous vehicles and be a leap toward artificial general intelligence that can operate in physical real-world situations. Instead of being based on hundreds of thousands of rules, algorithms, or lines of code like all previous versions of self-driving software, this new system had taught itself how to drive by processing billions of frames of video of how humans do it. Similarly, ChatGPT and other large language model chatbots train themselves to generate answers by processing billions of words of human text.

This means other industries like Health Care, Pharma and beyond become self-trained and self-organized rather than rule-based in a conventional sense. Instead of teaching a hospital the “golden standard” treatment path based on rules, AI and LLM will imitate experienced practitioners. Similar for shopfloor operators in Pharma, where resilience will be improved by imitating successful shopfloor as well as supply chain managers. Faced with a diagnosis, the neural network chooses a path based on what practitioners or operators have done in thousands of similar situations before, successfully. As much as it took Tesla to analyze millions of video clips on driving situations, Health Care and Pharma players or any other industry will need to fuel their systems with a huge number of quality proven examples of good treatment or effective operations. The pure mass of data facilitating self-learning will not only allow to imitate practices, however, there is a good chance that it will create new evidence, e.g., for an even more effective Health Care delivery as much as new practices to manage the Pharma supply chain and operations even more effectively in a “lot size 1” world.

Obviously, the speed of self-training and thus quality of AI and LLMOps depends on three gaps to be closed, at least. First, players with a high number of good-practice examples learn faster than others. Therefore, international players, e.g., established hospital chains as much as global Pharma market leaders, do have the chance to develop faster than others in their markets.

Secondly, relevant data must be mobilized by a proper infrastructure. For this reason, Cloud will help to gather and analyze data along the customer journey or value chain. In Health Care and Pharma for example, the cloudification will neither be limited to the electronic health record (EHR) of hospitals nor to data subject to serialization, material or network planning in Pharma, but will be fueled by comprehensive data models based on strategy-led architecture. S/4 HANA can play a significant role in here because ERP data are vital to success.

Thirdly, corporate data governance must evolve towards AI and LLMOps. In Health Care for example, scientific experiences about developing evidence through real-world data, instantly (i.e., over and beyond ex-ante clinical trials which provide real-world evidence) are rather pre-mature at this stage. Similarly, it is in other industries. Such processes need to be managed in line with guiding principles of Human AI to maintain the Chief Medical Officers` comfort feeling in hospitals in the future and leverage AI as well as LLM, boldly. In a same way, the Chief Operating Officers in Pharmaceuticals need to be able to determine the degree of resilience of their supply chain and be comfortable with GMP compliance. Of course, there are procedures to change procedures but, as of now, they are not necessarily AI- and LLM-compatible. Corporate data governance, therefore, is asking for a real change process to secure Autonomy and Auditing requirements in day-to-day business. For doing so, tech and business cultures must come together, appreciate each other and create a new, innovation-driven mindset that makes best use of AI.

Now that we understand what AI operations in terms of LLMOps are, let`s make a double klick on what Human AI operations looks like. To create benefit, AI must be run in a certain way. Based on good regulation for AI, which requirements must be considered when running AI systems from a corporate angle?

2.4 AI and LLMOps in line with Human AI

In addition to the four guiding principles from above, the fifth guiding principle is **Sovereignty**. Sovereignty means that any re-alignment or finetuning of LLM by end users or third parties must be quality assured by proper guidance. AI and LLMOps must refer to evidence and not just "please" users.

A latest study of scientists from the Princeton University and Virginia Tech shows that LLM models do not necessarily cover security risks when finetuning privileges are extended to end users, e.g., patients or less experienced health care professionals (for the study see Xianghu Qi et al. 2023). Finetuning means aligning the pre-trained LLMOps towards expected outcomes, such as better medicine. While existing security matching infrastructures can limit malicious behaviors of LLMOps at the time of inference, any unguided re-alignment by end users comes with a risk.

In Health Care for example, evidence-based medicine which refers to scientific studies must be the basis to improve quality of care delivery. Additional insights from AI and LLMOps based on real-time data can help, but also needs clear governance in line with guiding principles, and Autonomy in particular. To enable Autonomy, AI systems must be sovereign. Mechanisms like Reinforcement Learning from Human Feedback (RLHF) that foster alignment on end user level are critical. It can create outcomes that just focus on pleasing users, meeting their expectations or even show fake news in terms of hallucinations, rather than stick to evidence.

The sixth guiding principle is **Exchangeability**. As LLMOps builds on foundation models as a basis, Human AI is making sure these models can be replaced if needed, and thus foundation models do not take control. As much as Competition is needed to secure choices on a market regulation level, Exchangeability is a must on a corporate level. Exchangeability means to keep AI architecture foundation-model agnostic with the option to exit such models at any time. It means to be able at any time to transition the data and intelligence of LLMOps to a new foundation model, seamlessly. A foundation-model agnostic set-up of LLMOPs is key given the low number and transparency of relevant foundation models today. Practically this means, when running AI and LLMOps you really want to avoid a lock-in effect when contracting a foundation model provider.

The seventh guiding principle that supports Human AI operations and transparency is **Balanced-Open-Source**, i.e., sharing the source code of the AI and LLMOps to a certain extend. Open-sourcing basically can facilitate more humans to contribute to AI progress and enables large-scale collaboration. It can enable more expertise, more diverse views, and simply more human creativity and hours put into AI. This can drive innovation in new and beneficial downstream integrations, advance AI safety, and help develop AI capability. The more open we share source codes and provide access the more transparency we grant, in this respect open-sourcing is complementing and supporting Auditing. Open-sourcing of foundation models also can help to decentralize the influence over AI to a certain extent, away from major tech players by empowering smaller groups and independent developers. In this sense, open-sourcing can somewhat "democratize AI" by giving more humans influence over how AI is developed, optimized, and used.

But open-sourcing needs to be "balanced". Why that? The unguided access to highly capable AI for everybody can also create security risks at this stage. Decisions about open-sourcing of highly capable AI should be informed by rigorous risk assessments and managed or balanced, accordingly. Therefore, we talk Balanced-Open-Source. In the future, this can change as societal resilience to AI risk increases and improved safety mechanisms apply. Balanced-Open-Source also takes other ways to reduce corporate or autocratic control into account, like public participation, closed open-source communities, gradual model releases and model access for researchers (see also Seger et al. 2023).

Last not least, there is **Continuous Learning** as the eighth guiding principle. If you look back just two years from now, it is amazing how fast AI developed, how much we have learned about this technology and how little we have implemented from the overall AI potential. Cristal clear, the learning experience will go on and is asking all of us for a real extra mile in terms of adoption. Continuous Learning therefore asks for the ongoing effort to understand AI, go for opportunities as well as mitigate risks associated, sharpen our view as well as staying up to date on what Human AI is. Best-of-breed solutions and architecture must adopt, accordingly.

I need not say that the need to learn includes the guiding principles presented here. It would be foolhardy to believe that the requirements for Human AI presented in this essay are the last word in wisdom. This discussion mut be continued.

4. Our way ahead towards Human AI

The below figure is summarizing the guiding principles for Human AI. Good governance and operations that keep humans on the driver seat and make sure we benefit are vital to a successful development of AI.

No.	View	Guiding principle	Meaning
1	Governing AI	Independence	Governing and governed system to be separated to avoid conflicts of interest.
2		Autonomy	Final authority to articulate one`s own preferences and political will to stay with humans.
3		Auditing	IT assurance to secure transparency as well as legal compliance as well as the benefit of AI in terms of explainability, traceability, human oversight and continuous monitoring.
4		Competition	Antitrust to balance power and influence of AI systems and their providers. Industry policy to facilitate European champions.
5	Running AI	Sovereignty	No re-alignment or finetuning of LLM by end users or third parties without guidance. Adaptions to be quality assured.
6		Exchangeability	Keeping AI architecture foundation-model agnostic with the option to exit at any time as well as to transition to a new foundation model, seamlessly.
7		Balanced-Open-Source	Maintaining transparency through sharing the source-code or using other ways to reduce corporate or autocratic control, e.g., public participation.
8		Continuous improvement	Ongoing development of AI landscape to stay up-to-date and leverage best-of-breed. Including continuous development of guiding principles for Human AI.
9	Managing AI	Recursion	*- Subject to Part 2 -*
10		Empowerment	
11		Redundancy	
12		Self-reference	

Figure: Guiding principles for governing and running Human AI

While the question of good regulation and operations of AI have been answered in Part 1, we will focus on the good management of AI in Part 2:

How to manage digital technology with means of a human tech organization that fosters beneficial innovation? Digital technology and AI are creating so many expectations, some realistic, some futuristic. So, how can tech organizations deal with all this in a human way and how should CIOs set up to drive business innovation?

Imagine IT together with business experts feel like they had the right to design their own work. They were encouraged to influence and shape the tech-strategy, to define their own digitalization targets as well as use their own methods. They felt trusted to use their own judgment and to decide on IT investments and resources. They perceived themselves being on same-eyes-height with superiors and executives in the organization. As much they are being valued and appreciated as much they felt like “super-stars” in their organization who can make a difference. This is what “Human AI” is also about. Unfortunately, many IT organizations do not work like this. This is why we will do a double click on good AI management in the next part.

I am looking forward to catching up in Part 2 with you, appreciate your feedback as well as encourage all of us to stay excited, confident and alerted about the discussion on Human AI. Thank you!

Literature

Ashby, W. R., Introduction to Cybernetics. Chapman & Hall, 1956.

Chamorro-Premuzic, T., I, Human – AI, Automation and the Quest to Reclaim What Makes Us Unique, HBR Press 2023.

Daugherty, P. R.; Wilson, H. J., Human + Machine, Reimaging Work in the Age of AI, HBR Press 2018.

Freed, S., AI and Human Thought and Emotion, Taylor & Francis 2022.

von Goethe, W. J., Der Zauberlehrling. In: Goethes Werke. Gedichte und Epen I. Hamburger Ausgabe, Band I. C.H. Beck. München 1998, S. 276–279.

Isaacson, W., How Elon Musk set Tesla on a new course for self-driving, CNBC 2023.

Lee, T.; Trott, S., A jargon-free explanation of how AI large language models work - Want to really understand large language models? Here's a gentle primer., arstechnica 2023.

Qi, X., et al., Fine-Tuning Aligned Language Models Compromises Safety, Even when Users Do Not Intent To, - A Preprint (see also LinkedIn post of G. Wehberg sharing the study) 2023.

Russel, S., Human Compatible – AI and the Problem of Control, Penguin 2020.

Russel, S.; Norvig, P., Artificial Intelligence, A Modern Approach, 3rd edition, Pearson 2013.

Shneiderman, B., Human-Centered AI, Oxford University Press 2022.

Seger, E.; Dreksler, N.; Moulange, R.; Dardaman, E.; Schuett, J.; Wei, K.; et al, 'Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives', Centre for the Governance of AI, 2023.

Wehberg, G., Logistik-Controlling, Kern eines evolutionären Logistikmanagement, in: Jöstingmeier et al., Controlling-Konzepte im Wandel, Göttingen 1994, pg. 73 – 134.

Wehberg, G., Logistik 4.0 – Komplexität managen in Theorie und Praxis, Springer Gabler 2015.

Wehberg, G., Digital Supply Chains – Key Facilitator to Industry 4.0 and New Business Models leveraging S/4 HANA and Beyond, Routledge 2021.

Wehberg, G., Why LLMOps is the New Kid on the Health Care Block, LinkedIn post, 2023.