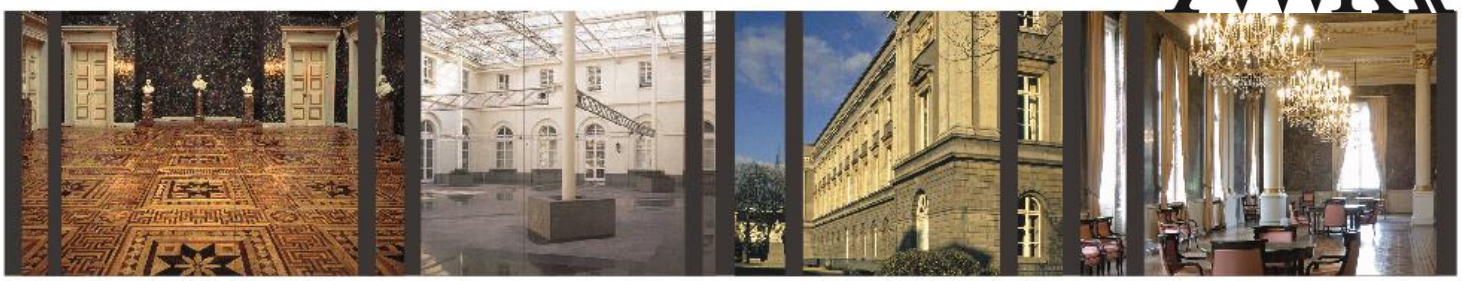
AI AS AN AGENT OF CHANGE

KVAB Thinkers Programme - 2023

Helga NOWOTNY

Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten
Paleis der Academiën – Hertogsstraat 1 – 1000 Brussel – België
info@kvab.be – www.kvab.be



Helga Nowotny

AI as an Agent of Change

Elizabeth Eisenstein's influential classic "The Printing Press as an Agent or Change" originally published in two volumes in 1980, was the trigger for the theme of the 2023 KVAB Thinker's cycle. It is an invitation to place AI into a larger historical frame – including the hype that surrounds it and the amazing efficiency that surprises even experts as well as the looming concerns to which it gives rise. Technologies do not fall from the sky, and it is wholesome, but also humbling, to reflect on the longer, often nonlinear, and unpredictable consequences that human inventions have generated in the societies in which they originated. The relationship between technology and society is never unidirectional. Technologies shape societies and their economies but are equally shaped by them. Societies adjust to the technologies that impact them in often unforeseen ways. They also appropriate them, inventing new and unplanned uses which may strengthen or undermine existing power structures, fulfill latent needs or, more generally, open the way to explore and exploit new opportunities.

A historical glance back: similarities and differences

'AI as an Agent of Change' is an intriguing metaphor. It places technological change intertwined with social change into a larger historical context, raising at least three questions which will be the guiding themes for this report. The first obvious question is about the similarities and differences, the continuities and discontinuities that can be found when comparing the societal impact of technological advances in AI/ML with those of preceding technologies. Undoubtedly, the printing press is a good start. Its impact was huge, for Europe and due to colonial expansion, far beyond. It induced changes that ranged from the proliferation of printers' workshops in European cities from 1500 onward to the disruptive effects it had on existing social and political structures. Ideas were disseminated through newly created social networks, leading to changes in mindsets which in turn greatly contributed to the rise of modern Science, the Reformation, and the European Enlightenment.

New markets for capitalist enterprises emerged, inducing further changes in how to finance them. The striking increase in the number of accessible books produced new audiences and readers. Book production correlated with the rise of diverse reading publics, initiating a long, albeit uneven, spread of literacy. In the words of Elizabeth Eisenstein "The fact that identical images, maps and diagrams could be viewed simultaneously by scattered readers constituted a kind of communication revolution in itself" (Eisenstein, 1980, p. 53). The result was a veritable knowledge explosion in the 16th century. Although this is often associated

with the discovery of the New World, access to a great variety of books and the ideas transmitted by them contributed at least as much. Galileo Galilei famously claimed that the Book of Nature is written in mathematics. The fact that Nature – and everything else that modern science continues to let us discover – is accessible to humanity owes much to the printing press and the societal changes it set into motion.

It is therefore tempting to draw parallels between the knowledge explosion of the 16th century and the 'information explosion' that holds us in its firm grip since some time. The recent public release of Generative AI based on LLMs, Large Language Models, has merely added to the overwhelming abundance of possibilities that AI/ML has opened. The convergence of computational power, the performance of neural networks and access to an enormous and growing amount of data has initiated the acceleration of most recent technological developments. It is another 'New World' that we are on the verge of discovering, the still largely unknown territory of 'digi-land' and what it holds in store for us. Many people fear that they are no longer able to cope with the speed and flood of information and what is demanded from them.

Social media – then and now

As often, drawing what at first appears to be obvious similarities with historical precedents quickly turns out to be more ambivalent upon closer inspection. Undoubtedly, the printing press opened new horizons for the reading audiences that avidly devoured whatever new knowledge and information could be obtained. This spurned the dissemination of ideas, leading to heated discussions and controversies that further propelled their dissemination. Today, we crave the new ever more. Social media are programmed to deepen this craving by targeting individuals or groups, leading them to retreat further into the bubbles of the like-minded. Our societies appear increasingly fragmented, and many blame the social media. Their recommending algorithms and ranking order reinforce preexisting tendencies, but they do so by engaging with users in specific ways. For instance, both the algorithms deployed by Facebook and the choices that users made, or were induced to make, played a non-negligible part in the 2022 US presidential election. The polarizing impact of FB algorithms is built into the content users get to see but so is what they chose to see. Devastatingly, users feed grows more polarized at every step of the recommending algorithm, leading users to engage more with the polarizing content (Uzogara, 2023).

This is only one of the many detailed, yet important mechanisms through which machines affect behaviour that remind us that machines are built to fulfill certain functions. They have human intentions inscribed into them. To be sure, propaganda was also rampant in the days of the printing press, when pamphlets and making fun of the authorities or slander attacks could be printed relatively cheaply and distributed quickly. But the difference when compared with the reach, speed and irreversibility of today's social media distribution is as obvious as worrisome. Fake news, we are told repeatedly, is nothing new, but at no time in history could Deep fakes be produced that make it virtually impossible to distinguish whether the face we see or voice we hear is genuine or not. The lines between 'true' and 'false' are becoming increasingly blurred, not only when it comes to statements about the real world, but also about its manifold digital representations. If the printing press was seen

as a threat to the religious and secular authorities of their time, today's digital technologies pose an enormous threat to the institutions and principles on which liberal democracies are built. Once the legitimizing distinction between 'true' and 'false' has been destroyed, we seem to be left with the arbitrariness of anomie or the submission to authoritarian rule.

Shifts in power – State versus corporations

Technologies initiate shifts in the structures of power. The printing press strengthened the centralization of power in the nation state. Printing helped to codify and standardize language and thus contributed to the rise of national identities. By contrast, a strong concentration of economic power occurs today in the hands of a few large international corporations, which governments and states are struggling to reign them in. They are at a loss how to protect citizens' rights and to deal with collective harm without strangling the potential of technological innovation. The challenges governments face range from the protection of privacy that citizens demand to whether enough new jobs will be created in time to replace those that will vanish. Nor is it known how a restructured labor market will affect one of the main pillars of the nation state, the system of taxation. Another looming issue to be tackled is connected to what a rapid diffusion of AI entails for the administration of public services, foremost the health and education system. In health care especially, data intensification and the integration of AI-assisted data practices entail a shift in control towards more standardization and greater efficiency, but also towards the private sector taking over many services now in the public realm.

Eisenstein reminds us that printing served the function of amplifying and reinforcing norms, values, beliefs, and ideologies. Today, we worry that the seemingly uncontrollable spread of fake news and conspiracy theories will further undermine what remains of common norms, values, and beliefs, creating a dangerous public void that can be filled by anything. As Hannah Arendt already warned some time ago against the rise of totalitarianism, once the world has become incomprehensible, people 'had reached the point where they would, at the same time, believe everything and nothing, think that everything was possible and that nothing was true' (Arendt, 1951). Such a situation lends itself, as we have seen during the pandemic, to an outright assault on the social authority of science-based expertise which, in the end, entails the abolition of the distinction between 'true' and 'false'.

Drawing historical similarities and differences therefore is never straightforward. We approach history through the lens of the most pressing concerns that occupy us in the present. The questions we pose are rooted and framed by what is foremost on our mind. History continues to be reinterpreted, partly because new sources and materials continue to emerge, but mostly because we pose new questions. Some arise from practical concerns and might guard us against the illusion that the latest technological wave is always 'revolutionary'. History is the best antidote against hype that has yet been invented. What we perceive as unprecedented, turns out to have precedents after all, even if they are only partial and highly selective. Nevertheless, we seek to learn how societies have coped previously with the challenges emanating from new technologies. What has worked, for the benefit of whom, and what have been the positive and negative effects seen with the benefit of hindsight?

AI – a General Purpose and System Technology

One such approach is offered by historians of innovation. There is general agreement that AI is a 'General Purpose Technology', GPT. This is an ensemble of technologies that have a wide range of applications across different economic sectors and industry. Their pervasiveness offers innovative complementarities, and their percolating effects tend to trickle down to lower levels. The long-term effects are therefore difficult to predict as it takes time until a systemic change that encompasses all sectors and levels of the economy has been achieved. It may also explain why we usually overestimate change in the short-term and underestimate it in the long term. The most prominent historical example of a GPT are electricity and electrification, including the role the down-sized small electric motor played in industrial production. The economic historian Carlotta Perez has analyzed the short- and long-term effects under the perspective of techno-economic paradigm changes. She shows that each of the previous major paradigm shifts has led to a quick concentration of wealth in the hands of a few entrepreneurs and of bold but ruthless investors and speculators. Sharp income gaps arise between winners and losers and a pervasive mentality of 'winner-takes all' dominates. In the end, governments had to step in, to ward off social unrest and/or to pursue a more solidary and progressive political vision (Perez, 2018).

Once a technology becomes mainstream, as is the case with AI applications in many fields and the rapid diffusion of Generative AI, change spills over and changes the economic ecosystem and its complex dynamics. Education, health, work, business will all be 'revolutionized' in the original sense of being 'turned over'. Such considerations are behind a conceptual approach that views AI as a 'system technology' which includes the wider technological and social ecosystem. It can then be compared with the effects that previous system technologies, the steam engine, electricity, the combustion engine, and the computer, have had. At a more pragmatic level, recommendation to the government about how to embed AI within society can then be derived from the history of previous system technologies (Prins, et al. 2021).

The historical look back allows to detect similarities and differences from which, hopefully, some lessons can be drawn for guidance. As 'lessons' from history always come with a big *caveat*, one of the more important take-aways messages for today is probably to sharpen the critical view of what *is* different this time. Obviously, this is not only the technology which brings amazing and significant advances compared to what was possible before. Rather, it enables us to see the larger picture in which technology is closely intertwined with society that absorbs, integrates, shapes, adjusts and appropriates in many different ways the technology it has generated. This happens through highly selective mechanisms, depending on existing social structures and practices which are mediated through complex processes. Based on shared practices, humans have the capacity to create new performative relationships, structures, and networks. The performativity arises from the use of symbols, from reinventing social relationships and from imagining collective futures. We call it culture – and we are active participants in an AI culture in the making.

Where are the citizens?

Another important aspect is the fact that technology cannot be separated from the power it confers. It can reinforce existing power structures or diminish them by enabling newcomers to gain power. Vested interests of the incumbents are always at play. Despite the rhetoric of innovation which dominates much of the official political discourse, the new is not always welcome and certainly not by those whose vested interests are threatened. During the early days of the internet a brief period prevailed which was infused by an emancipatory impulse. Many tech pioneers believed that the internet could exert a 'democratizing' influence, allowing everyone to participate and to share the benefits. Alas, such idealistic impulses were soon abandoned, greedily absorbed by what has become the Silicon Valley 'Tech Bro' culture, nurtured by its success and the belief – or the illusion – in its own illimitable power.

Recently, when ChatGPT was publicly released without asking anyone's consent, let alone considering the voices and needs of citizens, we became part of a large experiment conducted by OpenAI and its competitors. The struggle to regulate the power of the large international corporations has only begun and the attempts by technology insiders to introduce open source are in their infancy. Participation of citizens is reduced to the role of users in highly predefined and structured ways, following the operations of algorithms that have been designed to maximize 'clicks' and profit. The imbalance in financing AI research and development is glaring: only one tenth of investment in the US and in the EU comes from public sources, while the remaining 90% are private. This determines to a large degree also the directions of future research. The goal of turning AI into a public good is still far away.

Elizabeth Eisenstein's work is impressive because she takes a wider view of how society actively and selectively appropriated the opportunities the printing press offered. She shows how this invention was used by church and state, by capitalists, traders, and scholars, to suit and further their interests and beliefs. Technology can be used for different ends in different cultures; those in power can even suppress it, and attempts were made to do so. The interests of the elites, be they material or in the realm of ideas, always matter. Today, we find ourselves once more fully exposed to the different forces at work. The competition among the large corporations over market shares manifests itself in the bewildering variety of ChatGPT models that continue to be released, accompanied by the efforts of small start-ups that place their bet on open source in the hope to make a dent into the growing oligopolies of Big Tech. Obviously, the phase of consolidation has yet to set in. More worrisome are the geopolitical tensions between the USA and China. Among other, they are manifest in the fierce competition over the indispensable rare materials and the production of chips, resonating in Europe's call for a 'technological sovereignty'. The struggle over regulation in which the EU is the legislative forerunner with implementation as the difficult part to follow, has hardly begun. Attempts at reaching minimal standards for global regulation have still to be launched.

Thus, the comparison with the changes initiated by the printing press sharpens the critical view of the present situation. Despite some similarities, the differences are stark. And yet, as I will show, a continuity in the co-evolution between technology and humans can be

detected. It is a cultural co-evolution between humans and the machines built by them and, in biological co-evolution, it is open-ended.

Who is an agent of change and what is agency? The function of communication

The theme of the 2023 Thinkers cycle also poses the question of who is an agent and what is agency. The answers are far from obvious. Partly, because the definition of 'agent' varies enormously in academic disciplines, ranging from technical specificities in agent-based modeling to grand philosophical questions about free will. For pragmatic reasons, I prefer to occupy the middle ground, defining agency as the ability to actively interact with one's environment. Technology as an agent of change obviously is a metaphor.

We can start a long debate about who was the 'real' agent of change: was it the printing press as the forceful title of Eisenstein's book suggests or was there a multitude of agents of change, the numerous printers who set up their workshops in different European towns and those who financed them? What about the avid readers and the alliances or oppositions that formed between them and the ideas they sought to propagate? Moreover, the printing press could succeed only under specific institutional and cultural conditions to bring about the changes that followed. Woodblock printing in China dates to the 9th century and printing with moveable metal type was invented in Korea well before Gutenberg. It is obvious that a technology cannot be an 'agent' without the humans that invent, finance, operate, diffuse, and continue to improve it. A fortuitous combination of different actors and of cultural and institutional forces must combine with a technological innovation to generate the impact that the printing press achieved.

What distinguishes the printing press from other technologies is the function it assumed as a catalyst of communication. It is this function that served as a conduit for the dissemination of ideas, many of which were novel and subversive for the existing order. They were sufficiently appealing for the elites, and to those who aspired to become part of the elite, to adopt and use them for furthering their interests. The technology offered the means to reach the minds of people otherwise dispersed in far-away places, enabling to motivate and mobilize them. They all were agents of change, with differing interests and goals, yet united in making the best use of the technology according to their intentions. Communication became the means and the end at the same time, but – as always - the outcome remained unpredictable as it was open.

Communication as a catalyst for many pursuits is also a hallmark of 'AI as an Agent of Change'. Since the days of the invention of the printing press many new layers have been added to the function of communication. AI-based algorithms predict and are increasingly deployed in decision-making. But the basic idea of reaching other minds with specific content or messages, whoever and wherever they are, has persisted. AI/ML is capable to reach deeper into the cognitive and emotional state of users whose data are needed to target them as well as all the others with whom they are connected. Given enough data even those who do not use social media to communicate, can be identified. All these functions are attained by retrieving, storing, connecting, and processing information about the past of an individual, evidenced in the digital traces the user has left behind – which by

now means almost all of us. AI/ML has acquired impressive predictive power based on the extrapolation of these past traces and can combine them with information about all those with whom we have interacted in the past, generating a powerful tool for shaping the future.

The amount of data available for algorithms to be trained is staggering. To forestall the depletion of available data, recourse is already taken to create additional, synthetic data. AI/ML allows to build networks of networks, constituted by connections and interactions of various kinds. An enormous amount of information is thus accumulated about who we are, what we do, with whom, when and how we interact and how we feel. Thanks to sensors in cameras and satellites, installed above and below ground, AI/ML is capable to build a mirror world of the physical and social world we inhabit and enables interaction with it. Nearly every phenomenon and existing object by now is digitally documented or has a digital signature that can be followed, building new connections through iterations and almost infinite combinations.

The - relative – autonomy of machines: who controls agency?

We can conclude that AI *is* an agent of change, and yet, as with the printing press, it is an 'agent' only in the sense that we humans delegate and attribute agency to it. We let it perform for us, to attain goals set by us. We use it to come together and to set us apart. We delegate certain tasks to it, often oblivious of the consequences this might have. It becomes an extension of human capabilities, yet in doing so, we enter an ambivalent and open-ended relationship with a machine over which we do not have full control. We speak about 'complementarity' in carrying out certain tasks, but feel uneasy about the future, when the machines due to their amazingly efficient performance might take over ever more of what humans did before.

Automation will continue, this time replacing no longer physical labor, but increasingly cognitive tasks. The autonomy given to the machines is still relative. They depend on humans to supply them with the huge amounts of energy needed as well as for maintenance and repairs. They need infrastructures, including the organization to run the enterprise, investment strategies as well as legal and finance departments – the intricate hierarchies of the corporate world. Their further development still requires human brain power, and its numerous applications demand a skilled workforce, with continuous up-skilling and adapt at multi-tasking. But the overall direction clearly points to ceding more and more ground to digital machines.

Thus, a machine is nothing without the humans behind it. It is the artifact produced by humans that comes closest to what Nature has been doing throughout evolution – producing viruses that cannot replicate alone. A virus must infect a cell to make copies of itself. A machine needs human to keep it going and yet, as we observe with amazement, a digital machine can self-train and self-learn. The agency we have delegated to it, seems to extend ever further, raising serious questions whether we have delegated too much and in which domains and what needs to be done to maintain a kind of meta-control.

To inquire about agency therefore is a tricky task. It is usually defined as the ability of individuals to make their own decisions and take responsibility for their action. The sociological definition includes the power and resources of individuals to fulfill their potential. But can this or similar definitions of human agency be extended to machines and what do we mean when we transfer agency to an AI? In technical terms, machines are designed with various levels of autonomy, meaning that they have the ability to perform complex tasks with substantially reduced human intervention for specified periods of time and sometimes at remote distance.

In other words, an autonomous system is an agent or system (a machine or set of machines) that is driven and controlled to perform in accordance with the level of autonomy given to it. In practice, this can take on quite terrifying dimensions as is happening right now with the profound shift taking place in the militaries around the world, a shift towards AI, robotics, and autonomous warfare (The Economist, July 6th, 2023). It is no coincidence that a discussion recently broke out whether the UN Security Council should deliberate to set limits in the delegation of 'command and control systems' to autonomous weapons, akin to the non-proliferation treaties that were achieved to curb the spread of nuclear weapons.

The fear that humans might lose control over the machines they designed and built is not new and has existed since ages. Already Homer used the word 'automaton' ('acting of one's own will') to describe the automatic movement of wheeled tripods. Automated puppets that resemble humans or animals were used to demonstrate human ingenuity, to entertain and to deceive. The myth of *Frankenstein* lives on in innumerable manifestations. It has been revived in more civilized, yet also more insidious forms, in the Deep fakes produced by AI. In the guise of being more 'objective' than humans, it continues to be nurtured by the opaque operations of AI, the famous 'black box' algorithms. Technically and scientifically well-founded arguments have been brought forth to show that 'explainability' of AI is not possible (Lee, 2022). Even the best experts working at the forefront of Generative AI developments admit publicly that they do not (yet) understand fully the amazing performance accomplished by LLMs and that the question whether they produce 'emergence' remains open for the time being.

Whether AI will be able in the future to escape human control entirely and act completely on its own is one of the many speculations that the public is being fed to warn against a multitude of 'existential risks'. Situated in a faraway and hypothetical future, these risks pale however compared to those AI-powered battle ships without crews or self-directed drone swarms that are just two examples among the rapidly evolving technologies shaping the future of war right now. To have seen GPT-4 'showing sparks of artificial general intelligence' (Bubeck et al, 2023) or to state that Generative AI is about to 'develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control', as was written in the Open Letter, Pause Giant AI Experiment, of 29 March 2023, is an irresponsible use of hype that serves only to distract public discussion from the serious concerns and problems that need to be attended at present.

Our anthropomorphic tendencies

On a more mundane and practical level, humans in their interaction with artefacts have always attributed agency to them. This is deeply rooted in our anthropomorphic tendency to view the behavior of another entity or object in terms of mental properties. Daniel Dennett has told us how it works: "First you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in most instances yield a decision about what the agent ought to do; this is what you predict the agent will do" (Dennett, 1989, p. 17).

Apart from the philosopher's wording, this is indeed how we speak to the coffee machine or to the computer if it 'refuses' to do what we want it to do. We use anthropomorphic language every day in our interactions with machines. It is therefore not surprising that ChatGPT and its co-species that have been designed to communicate with humans induces us to say that it 'thinks', 'believes' or 'knows' – even if we understand that it is a non-thinking and non-believing, and certainly a non-sentient digital artefact that has 'only' been made to pretend that it thinks, understands, and believes. The unreflected use of such words in everyday language remains relatively harmless if it refers to familiar technologies that we have already incorporated into our world and hence learned to live with them. Yet it influences the ways in which we perceive the world. However, when it comes to AI it can transform the perception into a dangerously compelling illusion of being in the presence of a thinking creature like ourselves.

If unchecked and not critically reflected our anthropomorphic tendencies might turn against us and cause serious harm. This has been tragically highlighted by the suicide in Belgium of a man who engaged in week-long conversations with a 'therapeutic' AI (admittedly, an older generation than ChatGPT). Attributing agency to an AI program apparently contributed to the user's fatal decision. In my book 'In AI We Trust' I have highlighted the existence of a paradox that arises when we attribute agency to predictive algorithms and begin to believe that their predictions will come true. We leverage AI to increase our control over the future and uncertainty, while at the same time the performativity of AI, the power it has to make us act in ways it predicts, reduces our agency over the future. This happens when we forget that we humans have created the digital technologies to which we attribute agency. If unchecked, it might even bring about the return of a deterministic worldview in which most people believe that AI knows them better than they do themselves, including their future (Nowotny, 2021).

Social change: transitions and tipping points

The theme of 'AI as an agent of change' contains yet another question – how to understand social change. These days, we hear a lot about the various transitions we find ourselves in or which we should strive to achieve. The EU has programmatically proclaimed the 'twin transition' as going 'green' and 'digital'. Many governments have drawn up strategic

programs to achieve greater sustainability and how to manage the transition to get there. Yet, our knowledge of the processes that underlie societal change and may lead to a transition is rather poor. We can analyze them in retrospect and, for instance, identify some of the processes that lead up to tipping points. Numerous case studies of social change and of successful or failed innovation offer interesting findings, but the empirical evidence is usually confined to local cases. Often too small in size, too widely dispersed geographically and too divergent institutionally, these case studies hardly allow for comparability and generalization. On the macro scale, by contrast, simulations of complex adaptive systems based on mathematical tools and supplied with sufficient empirical data, can predict when and where in a complex network or system such tipping points are likely to occur. They are followed by transition or even collapse of the system. The gap between micro and macro remains and when it comes to understand societal change it seems that we are stuck between a rock and a hard place.

Yet here we are – in the middle of ongoing processes of societal change which will have enormous repercussions on individual lives and the future of our societies. Societal change has many dimensions, it is unequally distributed in its impact across different layers and sectors of a society. It is bound to produce winners and losers. Change is accompanied by promises and expectations, some of them deliberately overblown and others implicitly playing on latent needs or insatiable human desires. Promises are usually hard to keep and often end in disappointment. Expectations are to be carefully managed, a difficult task, as new technologies are usually surrounded by hype and tend to overpromise. In the more recent past, we have had our share: beginning with self-driving cars that were just around the corner; MOOCs that would 'revolutionize' the higher education system; the metaverse would soon take over our lives in the physical world and the promises of cryptocurrency luring many into reckless investments; not to speak about the fantasies of transhumanism that promise eternal life. The sign on the horizon of a brighter future remains the same: 'this time is different, just believe me'.

The long road ahead: AI as a public good

Yet, as the historical glance backwards reveals, this time *is* different – only we do not yet understand *how* and what it means. The experience of profound changes in our societies is ubiquitous and the turbulence linked to AI as an agent of change is as unsettling as is the prospect of a further acceleration of change. The predominant reaction so far has been the split between those who adopt techno-utopian visions and those who are immersed in their dystopian views. In a perverse way, this split feeds into and is fed by the already existing polarization in our societies, aggravated by the COVID-experience with its rise of the anti-vax movement and furthering distrust of citizens into their governments and experts. In addition, we are trapped by a dire outlook on climate change that no longer can be denied and surrounded by an economic recession that is about to begin. Geopolitical tensions keep rising while the war in Ukraine continues without prospect of a soon and good end. So, what is to be done?

A first step is to move away from the simplistic binary utopian-dystopian scheme of thought and to engage in a more sober assessment of risks and opportunities. These are not fixed

categories. Rather, they require a vigilant, flexible, and science-based understanding of what is at stake, for whom and under which circumstances. Maybe, the very concept of risk needs to be updated for AI as it no longer meets the simple definition inherited from the industrial age: probability of an event multiplied by the amount of damage. AI risk management and responsible AI practices are likely to become a key component in the future development of AI systems. Proper controls and taking context into account will be critical (National Institute of Standards and Technology, AI 100/1, 2023).

AI/ML is a powerful driving force of change, but it is not a force of nature to which societies and citizens are helplessly exposed. Despite many institutional flaws and the malfunctioning of existing institutions, our societies have sufficient means at their disposition to 'manage' risks, provided the political will is there. They can and must seize the opportunities that AI continues to open, even if it means to cope with challenges that will upset the existing order or overturn vested interest groups. In health care, for instance, AI/ML offers enormous opportunities for personalized predictive medicine (Hood and Price, 2023). Already now, it provides greater diagnostic accuracy and treatment options, with more rapid efficiency gains to come. If not carefully monitored, Big Tech is given access to data in return for AI-assisted services under contracts that may disadvantage the public health system, creating long-term dependencies under unfair conditions for the public health system.

Future historians will be able to reconstruct the outcomes which for us are unpredictable. What we – as scientists and as citizens – can do is to seize the opportunities of observing and analyzing the multiple processes of societal change in the making. We can gauge the leeways that exist to prevent harm and strike a reasonable balance between risks and opportunities. Above all, AI needs to be firmly institutionalized as a public good whose benefits should be available for all (Boulton, 2021). We can identify intervention points in the complex assemblage of AI/ML as a systems technology as well as in the finer technical and social details of its operations and recommend actions to be taken. Their chances of success will be enhanced if we can show the importance of bringing together governments, including the legislative branch; policymakers; industry; municipalities; media and the arts. Collectively, we need to create a renewed public space, a kind of 21st century *agora* recuperated from the occupation, if not obliteration, by social media and to promote its opening for a public discourse in which ordinary citizens are eager to participate.

Benefits from AI for society will only accrue if the terms of collecting, processing, and owning data and the delivery of services are not dictated by the large international corporations and the economic power they hold. Instead, it must be regulated by governments and include the participation and voices of citizens. AI must become a public good. The crass imbalance between private and public financing of AI research must be addressed, as it puts university-based research at a disadvantage regarding access to the needed computational power, data for training the algorithms, recruitment of talent and setting the directions of future research.

Finally, the call for a digital humanism with a human-centered focus in all AI-related technological developments only has a chance of being realized if a robust, institutionalized framework exists to back it up (Vienna Manifesto on Digital Humanism, 2019). Existing institutions were set up at another time to cope with a different set of problems. Time has

come to think earnestly about a new institutional framework that is better equipped and able to cope with many challenges that AI/ML brings, while laying the groundwork for exploring further and exploiting in a more equitable way the opportunities it offers.

AI and the outsourcing of knowledge operations

'One cannot not communicate' Paul Watzlawick, the communication theorist, famously declared, and we communicate all the time in many different forms. Some are analogue (with reference to an object) and others digital (logical and statistical connections). We communicate verbally, but also through body language. We transfer and exchange information, about ourselves, others, and the world. This can be ideas, practices, and knowledge at various levels of abstraction and complexity. Communication is a social practice which occurs in social settings. They can be symmetrical, at eye level and equal footing, or emphasize social hierarchies. Humans have developed elaborate codes that pervade all aspects of social life to distinguish themselves from others. Communication is at the root of the social organization of societies that has grown more complex over time.

Above all, it has stimulated and boosted the enormous growth of human knowledge as the result of the selective accumulation of the information that is communicated, enhanced and transmitted in multiple ways and means for multiple purposes. New ideas, knowledge or practices are combined, and recombined in novel ways whereby the content passes through selective filters in the processes of being transferred and exchanged. These filters are social and cultural. They follow the norms and values in a society that define which kind of exchange and content are culturally and socially valued and recognized. Societies rely on an explicit or implicit 'knowledge hierarchy'. For AI, the well-known knowledge pyramid, DIKW, shows different levels and seeks to explain the difference between AI as a knowledge-driven technology, while IT is data-driven. The pyramid's layers move upwards from data, to information, followed by knowledge and featuring wisdom at the top. In my book 'In AI We Trust' an entire chapter is devoted to wisdom needed in the future.

The technologies embedded in these knowledge hierarchies function to control which knowledge and information circulates. AI algorithms, like recommending systems and priority rankings, finetune these filter mechanisms further. Seemingly technical, they are designed to match the preferences, values and interests of the corporations that own them. The ongoing controversies between Big Tech and governments about whether enough is done by the former to contain or remove hate speech illustrates that who controls the media controls also the message, even more as the media have become the message, as McLuhan rightly diagnosed. The Catholic Church reserved the right to put books on the Index, whose content was deemed to go against its doctrine. Totalitarian regimes practice censorship while liberal democracies insist, in varying degrees, on the right of 'free speech'. Nevertheless, they too classify certain kinds of information as 'secret' whose diffusion might jeopardize national security interests.

The growing production of human knowledge

Evolution proceeds by variation and selection and a similar mechanism is at work in the growth of human knowledge. Selective filters operate not only to exclude, by controlling what is not to be communicated, but actively seek to include, absorb, and improve those communication that will produce new knowledge. The equalizing effect of printed editions, instead of fluctuating and unstable scribe products, was essential for the cumulative cognitive advance and incremental change that characterizes genuine scientific growth (Eisenstein, 1980, p.412).

The growth of human knowledge is greatly enhanced by technologies that enable the outsourcing, or externalization, of knowledge operations: processing and applying knowledge to other domains; storage and curation of data; dissemination of findings; novel combinations and the repurposing of knowledge. These and other operations, as well as the infrastructures and processes that underlie them, are essential for the selective uptake and the further reworking of knowledge through communication practices. Knowledge operations extend what is known in time and space, which would not be possible without outsourcing technologies. The history of humanity and what it was able to achieve so far is also a history of the outsourcing technologies deployed for the growth of knowledge.

Nowhere is this more evident than in modern science. One of its hallmarks was to make knowledge public and to share it, a radical break with the tradition of secrecy of knowledge-holders in previous times. By rendering the scientific findings and the processes how they were arrived visible and for all to see, new channels of communication were opened that greatly contributed to the spread of knowledge and the scientific world view. In doing so, science followed its own epistemic values while carefully delineating the boundaries over which it claimed cognitive and social authority. One of the epistemic values for developing and accessing scientific research underlies the practices of reproducibility, the theme of the Thinker's cycle 2022 (Leonelli and Lewandowsky, 2022). Science has excelled in optimizing its outsourcing practices. This is the reason why the scientific community very likely will succeed rapidly in harnessing the opportunities AI/ML offer, be it in drug discovery or literature-based discovery; numerical weather prediction, searching for new materials for batteries, designing new experiments or further automating labs.

The invention of writing as outsourcing a knowledge operation

AI as an agent of social change can therefore be seen as an integral part of the long trajectory of outsourcing knowledge operations with the help of technologies. It all began with the invention of writing which marked the transition of oral to written cultures. Writing was invented several times independently from each other, in different locations and at different times. It is an assemblage of constituent elements which includes the invention and mastery of symbols, like hieroglyphs, cuneiforms, and alphabets; the detailed elaboration of the physical substrates and infrastructures that were needed for the production, logistics, supply and use of adequate materials, like clay, stone, papyri, animal skins and others; the social competence and skills for collaboration and divisions of labor, like the specialization of scribes, the transmission of skills and of interpretative capabilities.

Taken together, these constituent elements form an assemblage that enabled communication to function more efficiently across time and space. Knowledge that previously would reside only in the memories of individuals and their oral communication skills (even if aided by mnemotechnic devices) and was orally transmitted from generation to generation, could now be outsourced and inscribed in a physical medium. An orator had the license (and often was expected) to modify the content in accordance with the occasion and the public addressed, while the words that had been inscribed in stone, on papyri rolls or on palm leaves created a temporal distance between the time when they had been written and when they were read and interpreted. Arguably, the new outsourcing practices also contributed to the capabilities of our ancestors for inventing and deploying abstract symbols giving rise to mathematics. The black (or white) board still used by mathematicians as the main medium to communicate with each other, supports this hypothesis.

The social and epistemic implications of writing were vast. For the first time, language was encoded in symbols that could be read, interpreted, understood, transmitted, and shared not only in novel ways, but deployed for a range of novel purposes. Measurements and numbers thrived and gained in importance. In the ancient world, Gods had their statutes and temples devoted to them, while writing was foremost deployed for taxation and trade. It was only with the rise of the monotheistic religions that the written word became the basis of sacred scriptures. Words could travel without a human pronouncing them. New networks of transmission emerged; trade became geographically extended and the measurement of the grain harvest to be taxed received a significant boost. For the first time, a direct confrontation with the past as fixed in writing ensued. This curtailed oral interpretative flexibility, but strengthened the weight given to the written word. Written contracts proved to be more reliable than oral ones, with further implications for trade, but also for peace negotiations.

As the sources were few and the material precious, control over them strengthened the centralization of interpretative authorities and led to a concentration of power in the hands of a small elite of priests, scribes, and rulers. Libraries became the repositories of all knowledge available, and their decline or destruction implied a significant loss of knowledge. Perhaps also for the first time, it became evident that a new technology was accompanied by the loss of certain cognitive facilities that humans had possessed earlier. As is well known, Plato deplored that the invention of writing brought with it the decline in the ability to memorize a vast corpus of knowledge.

What can the mechanisms and patterns that emerge in this first phase of the outsourcing of knowledge operations tell us? How does a social technology – writing – become an agent of change? There is no central, coordinating mechanism. As testified by the repeated times that writing was independently invented, human ingenuity is at work, producing symbols to communicate and to act through them. Mathematics as we know it is inconceivable without the writing of symbols. Outsourcing means that new spaces for communication and action are created, offering new opportunities while curtailing others. Some of these spaces will turn into ‘creative niches’, deploying the technology for yet to be invented purposes. As with every other technology, the uses and benefits of outsourcing knowledge operations are shaped by existing social and economic structures of power. In a highly skewed, unequal

society, the benefits will accrue disproportionally to those who have power. They will attempt to usurp the technology and use it not as an agent of change but to consolidate their power base.

And yet, the overall effect is one of expanding the knowledge base. Libraries became the physical storerooms, at first accessible only to the elite, but they remain the guardians of an important part of the human past, telling us what previous societies valued and how they saw and understood the world. Writing forms the basis for the sacred scriptures of the monotheistic religions until this day and it is difficult to imagine their influence without. Thus, outsourcing the word to a material substratum enabled words to detach from the local context in which they originated, transmitting, and exchanging knowledge with faraway places and with minds that eagerly received, contested, or appropriated them. However, the directions which the outsourced knowledge operations opened and the effects they produced, were impossible to predict.

From printing as outsourcing to social media

The second phase of outsourcing knowledge operations was initiated by printing which facilitated the exchange and diffusion of new ideas at an unprecedented speed and reach. New audiences and industries around publishing emerged. Outsourcing at a massive scale to books produced in large numbers enabled the revision and updating of older texts to incorporate new knowledge; to forge links among a readership widely scattered across Europe and enabled social movements to form and mobilize. It helped to spread literacy as the key to access the wider world out there and changed the attitude towards learning. It started a virtual circle, opening the way to be more inclusive and to foster participation. The advent of the printing technology coincided with the European discovery voyages around the world, fostering a greater openness towards a more cosmopolitan outlook which encouraged questioning and the spread of new ideas.

As detailed by Eisenstein, printing initiated a profound cultural change of mindsets, which ultimately marks this period as a crucial turning point in European history. The outsourcing of knowledge in books, newspapers, pamphlets, and illustrations meant that knowledge could no longer be monopolized by the elite but would reach a (relatively speaking) mass audience of those who were literate but whose numbers were growing. It had a major impact on the Renaissance with the revival of the classical literature; on the Protestant reformation as it enabled the interpretation of the Bible by each reader and thus shaped religious debates; on the Scientific Revolution as printing rendered possible the critical comparison of texts and illustrations; and by encouraging the rapid exchange of novel discoveries and experiments, giving rise to the Republic of Letters (Eisenstein, 1980).

It should be noted that some of the concerns we have today existed also during the cultural upheaval brought about by the printing press a few centuries ago. Religious and political pamphlets were full of hate and vile attacks on opponents (Darnton, 1984); fake news circulated widely, albeit much slower and more locally confined than today. The European Enlightenment had its dark side when it came to extending its claimed universalism to the colonies outside the Metropolitan area. The right to ‘free speech’ had still to become

constitutionally enshrined, while today, in a perverse twist, it is used in the US to argue for an almost limitless freedom to express racist and hate-filled opinions in the social media. Tellingly, in the emblematic confrontation between Church and Science, Galileo Galilei's trial was not about whether science was right or wrong. He had to abjure because he was accused to have violated the conditions the Church had imposed before allowing the publication of the 'Dialogue Concerning the Two Chief World Systems' in 1632.

The profound transition we experience today, triggered by the amazing advances in AI/ML, concords with the evolution of outsourcing of knowledge operations of previous phases. Yet its effects will be orders of magnitude larger. Outsourcing is no longer limited to inscribing words on material and make them travel across time, nor to disseminate ideas through cheap paper to newly created audiences. Considering the time scales covered by the previous phases, the information and communication technologies of the late 19th century and 20th century, telephone and telegraph, radio and TV, function merely as a prelude for today. They inaugurated the shrinking of distance around the world, while increasing awareness of what happened elsewhere. The mass media introduced one-to-many communication, followed by many-to-many communication, individual targeting, and user-generated content once the Internet took over, followed by the ubiquitous spread of social media.

Generative AI: the outsourcing of knowledge production

The big jump in outsourcing knowledge operations based on LLMs consists in the fact that the production of knowledge itself is outsourced. By training, and teaching self-training, to ever more sophisticated algorithms with trillions of tokens, consisting of all texts, images and sounds available on the internet, humans have delegated the production of new knowledge to the machines designed and built by them. Although 'only' extrapolated from the past and based on probabilities, the combination results in generating something new. Whether the answers are correct or made-up, factful or hallucinations, is another matter to be critically assessed. If automation run by AI consists in outsourcing hard or tedious physical tasks from humans to machines, Generative AI takes over an increasing number and range of cognitive tasks outsourced to it. ChatGPT is designed as dialogue with a digital Other and it is through dialogue – the questions asked, the prompt engineering that is undertaken – that new knowledge results. Given that outsourcing began with a shift from an oral to a written culture, it is an ironic twist of history that Generative AI signals a partially return to an oral culture. It becomes important again to know how to dialogue and have a conversation, this time with a machine.

The outsourcing of knowledge production to digital machines brings a series of challenges with it and some of the most pressing ones will be dealt with later in this Report. The advantages of this last and most radical step in outsourcing are huge, and their integration into our individual lives and the functioning of our societies carry explosive potential. For example, AI/ML is already used to find the most promising prescription 'cocktail' of medication for the precise treatment of specific, rare types of cancer. In doing so, it outperforms the most experienced doctor, as it has access to a trove of the latest medical literature. This raises the fundamental question of how doctors will be trained in the future.

Will they become supervisors of the AI? Perhaps. Similar questions crop up in many other fields of application where the benefits are obvious, but the role of humans becomes ever more elusive and in urgent need to be redefined.

Perhaps the greatest, unintended and undervalued gift by Generative AI is that it opens a range of fascinating new research questions. They range from in-depth explorations how the human brain works in solving tasks compared to that of an AI; to questions about the future evolution of language once LLMs have become ubiquitous in daily life; the impact of ever more intimate and intense interactions with AI, especially on the younger generation and the formation of identity; to questions about the impact of AI on liberal democracies and what can be done to stop further erosion.

Beyond such research questions and the launch of new research fields, science has an important role to play in conveying to the public how it works. The physicist Richard Feynman once said, 'Science is what we have learned about how to keep from fooling ourselves'. In view of the design of ChatGPT to make believe one communicates with a human and given our anthropomorphic tendencies, it is even more important for science to bring Feynman's insight to the public. The pandemic made painfully clear how little politicians and the public understand that science is organized skepticism and that to question claims about scientific findings in an elaborate process of verification and validation, is an essential epistemic virtue of science and not a fault.

Hence, to exemplify in understandable and accessible terms how to think in a critical, yet constructive way when dealing with AI/ML is one of the main responsibilities that falls upon scientists. How does a scientist respond when people tell her that 'AI knows me better than I do myself' and when they start to believe that AI is an agent whose predictions will inevitably turn out to be true? The tacit assumption in science communication still is that once a certain level of digital literacy is achieved, citizens would act rationally and adopt digital solutions and the behavioral recommendations that come with them. But it does not work like that. Empirical research has proven that we need to move away from the 'defict model' of science communication which attributes refusal to accept what scientists say as lack of understanding (Wynne, 1993). Instead, to engage in accessible terms with the public entails to 'show and tell' how science finds productive ways of making AI support its pursuits. Scientists are in a unique position as they already use AI widely to assist in their research. They can show concrete examples and the advantages derived from it, be it in medical, environmental, or other fields of research. At the same time, they must communicate how it works so that 'science keeps us from fooling ourselves'.

In this Report I have laid out the tapestry, based on my observations and analysis of 'AI as an agent of societal change'. After inspiring and intense discussion with stakeholder groups and the Steering Committee of KVAB, we agreed on the following actionable recommendations.

Recommendation 1:

We recommend launching a broad public campaign under the provisional motto "AI for citizens – citizens for AI" to support citizens to appropriate and use AI for their benefit and a better society.

The aim is to deepen and spread the understanding of how AI and digital systems work, to explore the potential of current and future applications,
their use and to learn about their limitations.

The many already existing and emerging initiatives should be given the official mandate to

(1) coordinate amongst themselves the educational efforts directed towards these goals;
(2) specify and map their respective target groups (age groups, formal and informal settings, etc.), the means and materials they use, test and develop (e.g. for teachers in primary and secondary schools), forms of cooperation with universities, media, the arts and industry;
(3) create ample space for continuous exchange of experience and mutual learning across academic disciplines and generations;
(4) ensure that all educational efforts include a digital humanism perspective (and therefore go far beyond digital literacy) https://informatics.tuwien.ac.at/digital-humanism/

Towards this end, a robust institutional framework should be established and provided with the necessary financial and personnel resources, initially for a period of three years, renewable after evaluation.

Recommendation 2:

We recommend making basic research in AI a high priority to be carried out in an ERC-like mode (bottom-up, PI-centered). This would counteract the dominance of a one-dimensional 'technological solutionism' that ignores and/or sidelines alternatives in the choice of research problems, methods, and techniques. It should include a more humanistic understanding of the range and depth of human experience and what it means to be human.

The present overconcentration of financing AI-related R&D in the private sector generates a worrisome imbalance for (mainly) university-based independent research regarding access to computational power, training data, attracting talent, and pioneering new directions of research. In the interest of AI as a public good, these disadvantages must be addressed.

The field of AI, including ML and Generative AI, is relatively young and lacks a historical perspective, especially in Europe. This entails the loss of valuable technical know-how, mathematical concepts, techniques, and scientific insights. Promising lines of research were

often prematurely closed. Only a strong focus on basic research can initiate their rediscovery and further exploration of historical paths that were not taken.

Recommendation 3:

We recommend a vigorous support of research on the impact AI has on society regarding aspects and in areas unlikely to be taken up by the large international corporations.

As we are only at the beginning to systematically follow and analyze the possible beneficial applications of AI for different groups in society and to learn about the avoidance of social harm, it is crucial to include the rapidly evolving experience, voices and needs of citizens.

Students of AI and related technical fields (and their teachers) should be encouraged to include a digital humanism perspective in their technical training and practice. Likewise, students in the humanities and social sciences (and their teachers) have to become more familiar with the technical aspects.
These are the preconditions for more and better grounded inter-, and even trans-disciplinarity, that is urgently needed.

References:

Arendt, H. (1951) The Origins of Totalitarianism. Berlin: Schocken Books.

Boulton, G.S. (2021). Science as a Global Public Good. *International Science Council Position Paper*. https://council.science/wp-content/uploads/2020/06/Science-as-a-global-public-good_v041021.pdf

Bubeck, S. et al. (2023) Sparks of Artificial General Intelligence: Early Experiments with GPT-4, (24.3.2023), https://arxiv.org/abs/2303.12712.

Darnton, R. (1984) The Great Cat Massacre and Other Episodes in French Cultural History. New York City: Basic Books.

Dennett, D.C. (1989) The Intentional Stance. Cambridge: The MIT Press.

Eisenstein, E.L. (1980) The Printing Press as an Agent of Change. Cambridge: Cambridge University Press.

Hood, L. and Price, N. (2023) The Age of Scientific Wellness. Why the Future of Medicine is Personalized, Predictive, Data-Rich, and in Your Hands. Harvard University Press: Cambridge, Mass.

Lee, E. A. (2022) Limits of Machines, Limits of Humans. DigHum Lecture, https://caiml.dbai.tuwien.ac.at/dighum/dighum-lectures/edward-lee-limits-of-machines-limits-of-humans-2022-05-24/.

Leonelli, S. and Lewandowsky, S. (2022) The Reproducibility of research in Flanders: Fact finding and recommendations. KVAB Thinkers' report 2022.

National Institute of Standards and Technology (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0): https:doi.org/10.6028/NIST.AI.100-1

Nowotny, H. (2021) In AI We Trust. Power, Illusion and Control of Predictive Algorithms. Cambridge, UK: Polity Press.

Perez, C. (2018) Second Machine Age or Fifth Technological Revolution? (Part 4) The Historical Patterns of Bounty and Spread. (21.11.2018). https://medium.com/iipp-blog/second-machine-age-or-fifth-technological-revolution-part-4-4420c29ceed.

Prins, C., Sheikh, H., Schrijvers, E., de Jong, E. and Steijns, M. (2021), Mission AI. The New System Technology. Summary of the Dutch report Opgave ai. De nieuwe systeemtechnologie published by the Netherlands Scientific Council for Government Policy www.wrr.nl.

The Economist, (2023) A new era of high-tech war has begun, The Future of War. (06.07. 2023). https://www.economist.com/leaders/2023/07/06/a-new-era-of-high-tech-war-has-begun.

Vienna Manifesto on Digital Humanism, (2019) https://caiml.dbai.tuwien.ac.at/dighum/dighum-manifesto/

Wynne, B. (1993) Public uptake of science: a case for institutional reflexivity. *Public Understanding of Science,* 2 (4), pp.321-337.

Uzogara, E. (2023) Democracy Intercepted. Did platform feeds sow the seeds of deep divisions during the 2020 US presidential election? *Science* Vol. 381 Issue 6656, 28 July 2023, pp.386-387.