AI AS AN AGENT OF CHANGE

KVAB THINKERS' REPORT 2023

KVAB Press

KVAB POSITION PAPERS
85 b

Cover concept: Francis Strauven
Cover design: Charlotte Dua
Image: Shutterstock

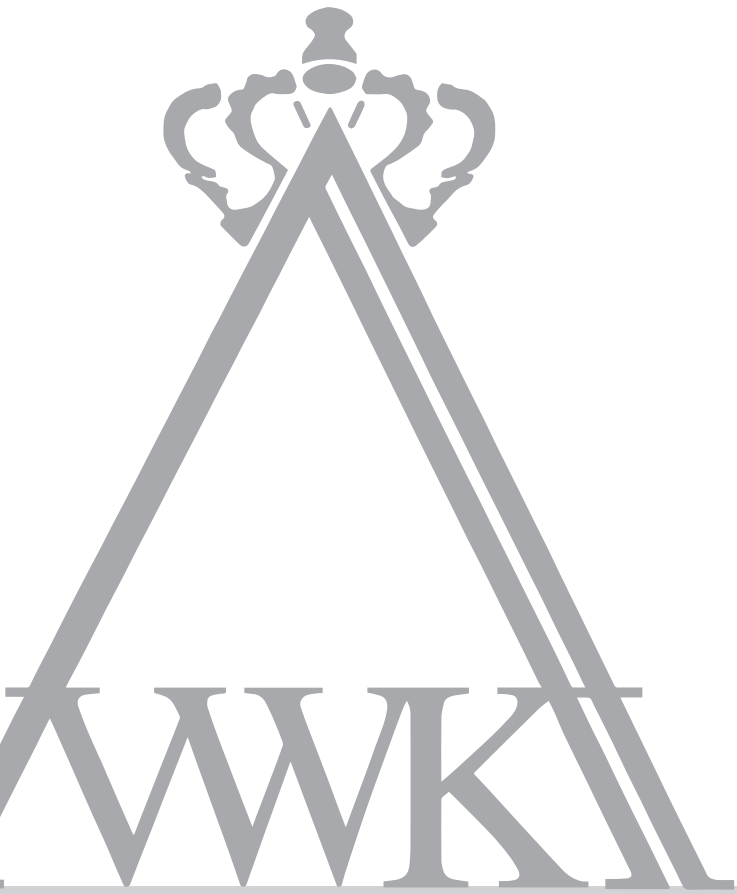
The drawing of the Palace of the Academies is a reproduction of the original perspective made by Charles Vander Straeten in 1823. The logo of the KVAB was designed in 1947 by Jozef Cantré.

The KVAB Position Papers (Standpunten) are published by the Royal Flemish Academy of Belgium for Science and the Arts, Hertogsstraat 1, 1000 Brussel.
Tel. 00 32 2 550 23 23 – info@kvab.be – www.kvab.be

AI AS AN AGENT OF CHANGE

KVAB THINKERS' REPORT 2023

Helga Nowotny
Ine Van Hoyweghen
Joos Vandewalle



D/2023/0455/08
ISBN 9789065692269

Recommended citation: Helga Nowotny, Ine Van Hoyweghen,
Joos Vandewalle, *AI as an agent of change*, *KVAB Thinkers' report 2023*,
KVAB Standpunt 85b, 2023.

Printing office Universa Press

AI AS AN AGENT OF CHANGE

TABLE OF CONTENTS

Executive Summary

This position paper was produced in the context of the Thinker's Program of the KVAB. Extensive consultations with stakeholders provided the Thinker-in-residence with insights about current practices and possible futures in Flemish AI research and society at large.

Even though Artificial Intelligence (AI) tools are not magic but the simple result of mathematical optimization, massive computer power and data sets, these tools are imbued with extraordinary promise. They have managed to generate substantial benefits already, particularly in the form of enhanced efficiency, accuracy, timeliness and convenience across a wide range of research and services, products and processes, including healthcare, genomics, fusion research, product design and simulation, autonomous systems, predictive maintenance, quality control, inspection processes and environmental impact assessment.

At the same time, the emergence of AI has been accompanied by rising concerns about its potential risks and damaging effects: for individuals, for vulnerable groups and for society more generally. This raises questions as to how AI will affect our everyday life, in both its private and its professional contexts, as well as influence our views of who we are as human beings. This requires a human-centered approach in the design, use and further development of AI, which entails an alignment with human values and needs. We must shape technologies in accordance with human values and needs, instead of allowing technologies to shape humans. All of this is growing more urgent by the day. As generative AI systems (e.g., ChatGPT) develop at lightning speed, the scientific community must take a central role in shaping the future of a human-centered approach to AI.

The aim of the KVAB's Thinker's Cycle was to reflect on the impact of these latest developments in AI, as well as on the question of its implications for our view of humanity: autonomy, human agency and the need for a "digital humanism." The title of the Thinker's Cycle, "AI as an Agent of Change," was inspired by the two-volume work of the historian Elisabeth Eisenstein, *The Printing Press as an Agent of Change* (1980).

The consultations addressed the latest developments by providing a broad philosophical, sociological and historical perspective on the impact of AI. Drawing on the specific backgrounds, perspectives and competencies, this position paper presents not only a consistent analysis, but also valuable approaches and proposals for taking further steps. Given the quality of the Thinker's discussions with participants and her constructive findings, these efforts provide a solid foundation for the productive integration of AI into science and society in Flanders.

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.

Toward this end, a robust institutional framework should be established, including allocation of the necessary financial and human resources, initially for a period of three years, and potentially 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 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 the more and better grounded interdisciplinarity, and even trans-disciplinarity, that is urgently needed.

Preface

Artificial Intelligence
human company
invisible algorithms
future needs wisdom

(Nowotny, 2021)

Each year the Royal Flemish Academy of Belgium for Science and the Arts (KVAB) organizes, within the Thinker' s Program, two so-called Thinker's Cycles at the initiative of one of its Classes and/or Reflection Groups. In each Cycle, the Thinker(s) invited is introduced to the specifics of a particular societal challenge in Flanders. These Cycles result in position papers whose views and findings may be included in policy recommendations prepared by the European Federation of Academies in the context of SAPEA Science Advice for Policy by the European Academies (https://www.sapea.info).

In 2022, the KVAB Class of Humanities (KMW) proposed the Thinker's Cycle on "AI as an Agent of Change," to address the impact of Artificial Intelligence (AI) on science and society. A Steering Committee of this Thinker's Cycle selected the international expert Prof. dr. Helga Nowotny as Thinker-in-residence, because of her extensive expertise on this topic in relation to academia and science policy. In consultation with the Steering Committee, she engaged in debate with relevant experts, stakeholders and practitioners in Flanders. On the basis of these interactions, she developed her independent assessment. As Thinker-in-residence she presented her findings and recommendations at a final, public symposium.

It is our pleasure to thank all those who contributed to the success of this Cycle: the Steering Committee, the consulted experts and stakeholders, the public and the KVAB staff. In particular we would like to thank our Thinker-in-residence, Helga Nowotny, for an excellent report and her recommendations. She formulated her long-term vision on the impact of AI in a thought-provoking way, while integrating the information she derived from her interactions with Flemish experts and stakeholders. As such she has provided an excellent basis for the Flemish scientific community to expand the discussion on integrating AI in science and society.

The present publication is a position paper based on the findings and recommendations of this Thinker's Cycle. This paper was digitally approved for publication by the KVAB Class of Humanities (KMW) on December 15, 2023.

Ine Van Hoyweghen
Coordinator of the Thinker's Cycle
November 23, 2023

1. Introduction: Positioning, Aim and Approach of the Thinker's Cycle

Ine Van Hoyweghen

The capabilities of Artificial Intelligence (AI) systems have grown quickly over the last decade. Employing a growing wealth of algorithmic insights, access to massive data sources and computational power, AI researchers have created systems that can comprehend language, recognize and generate images and video, write computer programs and engage in scientific reasoning. If current trends in AI capabilities continue, AI systems could have transformative impacts on science and society. Powerful AI systems will come with significant benefits and risks. This raises questions as to how AI will affect our everyday life, in both its private and its professional contexts, as well as influence our views of who we are as human beings. It is within these current and ongoing transformations that the Thinker's Cycle was developed, with the principal aim of reflection and debate on "AI as an Agent of Change."

Positioning of the Cycle

In recent years, AI has attracted a great deal of attention in industry, education, research, politics, government and society at large. The rapid advancements in artificial intelligence (AI) have led to the development of generative AI, including large language models (LLMs) that can produce fluent outputs such as text, images and code on the basis of the patterns in their training data. ChatGPT for example is a large language model (LLM), a machine-learning (ML) system that autonomously learns from data and can produce sophisticated and seemingly intelligent writing after training based on a massive dataset of text. It is the latest in a series of such models released by OpenAI, funded largely by Microsoft, while other major tech firms are racing to release similar tools (van Dis et al., 2023).

Even though Artificial Intelligence tools are not magic but the simple result of mathematical optimization, massive computer power and massive data sets, these tools are imbued with extraordinary promise. They have managed to generate substantial benefits already, particularly in the form of enhanced efficiency, accuracy, timeliness and convenience across a wide range of research and services, including healthcare, genomics, fusion research, product design and simulation, autonomous systems, predictive maintenance, quality control, inspection processes and environmental impact assessment. AI tools are becoming increasingly common in science, and many scientists anticipate that they will soon be central to the practice of research, as surveyed in a recent *Nature* paper involving more than 1,600 researchers around the world (Van Noorden & Perkel, 2023). Scientists have used these models to help summarize and write research papers, brainstorm ideas and write code, while others have

examined the potential of generative AI to help produce new protein structures, improve weather forecasts and suggest medical diagnoses, among many other ideas. According to the OECD (2023), accelerating the productivity of research could in fact be the most economically and socially valuable of all the uses of artificial intelligence (AI).

At the same time, however, the emergence of AI has been accompanied by rising concerns about their potential risks and damaging effects: for individuals, for vulnerable groups and for society more generally. Recently, we have seen tech industry insiders trumpeting the "existential risks" of artificial intelligence (*Nature* Editorial, 2023). Various open letters were published with thousands of signatures advocating a pause in training AI systems more powerful than GPT-4. Unchecked, AI development "might eventually outnumber, outsmart, obsolete and replace us," or even cause "loss of control of our civilization," one of the letters warned. But as critiques of these letters point out (see, e.g., *Nature* Editorial, 2023), this focus on hypothetical risks ignores actual social harms and risks happening already. AI is reinforcing and exacerbating many challenges of today's world, such as bias, unwanted profiling/discrimination, disinformation, data misuse, closed access and rising social inequalities.

These risks are (and have been) well documented by scholars in the social sciences and the humanities (see, e.g., Crawford, 2021; Benjamin, 2019). Many of the machine learning (ML) models are black boxes that do not explain their predictions in a way that humans can understand. These black-box machine learning models are used for high-stakes decision-making throughout society, causing problems of bias and discrimination in healthcare, social policy, insurance, and other domains (Obermeyer et al., 2019; Rudin, 2019). Biased AI systems could use opaque algorithms to deny people welfare benefits, medical care, or asylum — applications of the technology that are likely to most affect those in marginalized communities (Kalluri, 2020). One of the biggest concerns surrounding the latest breed of generative AI is its potential to boost misinformation and "deep fakes" – videos of synthetic faces and voices that can be indistinguishable from those of real people. In the long run, such harms could erode trust between people, politicians, the media and science, especially in the absence of rules on the production of the underlying models and codes (Jones, 2023; van Dis et al., 2023; Van Noorden & Perkel, 2023). The underlying training-sets and LLMs for ChatGPT and its predecessors are not publicly available, and tech companies tend to conceal the inner workings of their generative AIs (Ferrari et al., 2023; Bockting et al., 2023).

All of this calls for a well-balanced governance approach – one that respects societal values and is publicly supported before the technology undermines science and public trust. This concern was highlighted at the European policy level. In her State of the Union in September 2023, Ursula von der Leyen, President of the European Commission, called for a global approach to understanding the impact

of AI, modeled on the Intergovernmental Panel on Climate Change (IPCC), with a brief to "set minimum global standards" for safe and ethical use of AI (EC, 2023). This new body on the benefits and risks of AI for humanity will consist of scientists, tech companies and independent experts. The call for "responsible AI" is in line with the ground-breaking legislation the Commission proposed in April 2021, the AI Act (EC, 2021), which is currently being negotiated by MEPs and member states. The Act, which imposes market rules on AI-powered systems according to their potential risks for society, is considered "already a blueprint for the whole world" (EC, 2023). While the EU is in the process of finalizing its first regulation on artificial intelligence, the scientific community has yet to come up with a unified response on how generative AI could be used in higher education and research. Plans are underway to set up a dedicated new unit at the Commission's research directorate to lay down guidelines, as well as for a debate on how to handle the use of AI in science to be initiated as part of the European Research Area (ERA) policy agenda. In July 2023, the Commission's science advisors published a scoping paper on the issues involved, pointing to a lack of "dedicated and systemic policy facilitating the uptake of AI in science" (SAM, 2023).

The impact of AI is a multifaceted theme with many angles and areas, as well as multiple stakeholders and policy levels. Faced with this latest technological change, people instinctively turn to technologists for solutions. But the impacts of AI cannot be mitigated through technical means alone; solutions that do not include broader societal insight will only compound AI's dangers (Lazar & Nelson, 2023). This requires a human-centered approach in the design, use and further development of AI which entails an alignment with human values and needs. We must shape technologies in accordance with human values and needs, instead of allowing technologies to shape humans. All of this is growing more urgent by the day. As generative AI systems develop at lightning speed, the scientific community must take a central role in shaping the future of a human-centered approach to AI.

The Cycle's Aim and the Thinker

The aim of the KVAB's Thinker's Cycle was to reflect on the impact of these latest developments in AI, as well as on the question of its implications for our view of humanity: autonomy, human agency and the need for a "digital humanism." The title of the Thinker's Cycle, "AI as an Agent of Change," was inspired by the two-volume work of the historian Elizabeth Eisenstein*, The Printing Press as an Agent of Change* (1980). The advent of printing technology was, quite literally, an epoch-making event. The shift from script to print technologies revolutionized Western culture. In her book, Eisenstein argues that the revolutionary transition from a culture of manuscripts to a culture of print had a fundamental influence on the Renaissance, the Protestant Reformation and the rise of the Scientific Revolution. The age of AI now dawning may resemble this transformation in that it is likely to

generate not only myriad benefits, but also unintended effects not recognized at the time of the technology's unfolding.

To navigate these questions of the Cycle, we were honored to have Helga Nowotny as our Thinker-in-residence. Helga Nowotny is Professor Emerita in Science and Technology Studies at EHT Zurich. She is a Founding Member and a Former President of the European Research Council (ERC). She is a foreign member of the Class of the Humanities (KMW) of the KVAB (see Annex 1 for CV). She has closely followed developments in AI for more than half a century. In her book *In AI We Trust* (2021), Nowotny addressed the latest developments by providing a broad philosophical, sociological and historical perspective on the impact of AI. She thereby points to an inherent paradox of our trust in AI: "We want to use AI to better control our future but through its predictive algorithms, AI reduces our freedom to shape such a future. AI must therefore be flanked by our human capacity as an 'agent of change' to maintain a shared, open future" (Nowotny, 2021).

The interests and concerns of the Thinker's Cycle align with those of the Digital Humanism Initiative at the Vienna Institut für die Wissenschaften vom Menschen (Vienna Manifesto on Digital Humanism, 2019). This international collaboration seeks to build a community of scholars, policymakers and industrial players who are focused on ensuring that this technology development remains centered on human interests. Digital humanism observes and describes digital technology changes and aims to shape and influence the development of these technologies and policies towards the values of human rights, democracy, participation, inclusion and diversity. Similar initiatives were set up globally in recent years. In 2019, for example, the Institute for Human-Centered AI (HAI) was established at Stanford University. Publications and thematic meetings underscore the urgency of the concerns involved, including freedom, algorithmic transparency, cognitive sovereignty and "hybrid mind" in human-machine symbiosis.

The topic of AI is currently much debated in Flanders as well, appearing in almost every issue of the biweekly announcements of the Flemish Advisory Council for Innovation and Entrepreneurship (VARIO). AI is also the subject of Flemish and nationwide Belgian policy (Flemish Policy Plan for Artificial Intelligence (VAIA, 2019)) and the National Convergence Plan for the development of Artificial Intelligence (Federal Public Service, Policy and Support (BOSA, 2022)). Several KVAB Thinkers' Cycles have focused on the theme of AI and digitalization and were reported in the series of KVAB Position Papers: "Artificial Intelligence, towards a fourth industrial revolution" (Steels et al., 2017), "Societal values in digital innovation: who, what and how?" (Rabaey et al., 2019) and a recent KVAB-ARB joint position paper "A call for an accelerated digital transformation for Belgium" (Vandewalle et al., 2022). Other academies and international umbrella organizations pursue debates and activities that are also important for the Thinker's Cycle. ALLEA, for example,

whose code of conduct for research integrity is one of the guiding documents for Horizon Europe, updated this framework earlier this year to reflect the changes brought about by AI (ALLEA, 2023).

In collaboration with the Thinker-in-residence, the aim of this Thinker's Cycle was to examine on the basis of a broad reflection, proposals for supplementary policies that are developed on a local, national and international level to stimulate the productive integration of AI in science and society. The Cycle provides contributions and advice regarding the objectives proposed by the Flemish government for 1. strategic basic research, 2. training needs, 3. ethical challenges, and 4. outreach to the general public.

Approach

The Cycle was proposed by the KVAB Class of Humanities (KMW), as an initiative of members Marc de Mey and Ine Van Hoyweghen. The proposal was accepted by the Class of Humanities on November 19, 2022, after which a starting note and call for participation to the Steering Committee was distributed to all KVAB members. In December 2022, a Steering Committee was put together, composed of members from the different Classes of the KVAB, the Young Academy and other relevant experts (see Annex 2). The role of the Steering Committee was to ensure proper underpinning of the Thinker's activities and to provide necessary input. The Thinker-in-residence was given ample freedom and remained completely independent in writing the report and its recommendations. By working together with the Steering Committee and numerous Flemish experts and stakeholders, she managed to make a significant contribution to the topic of the impact of AI by articulating a long-term vision and, in this way, contribute to Flemish policymaking. She developed her views and recommendations after several rounds of intensive meetings and consultations with experts and stakeholders across Flanders.

The experts, stakeholders and practitioners contributed to these discussions. They typically come from all the relevant institutions like

- KU Leuven
- VUB
- University of Antwerp
- University of Hasselt
- Research Centers (representatives from imec, Flanders AI, VIB, …)
- Umbrella organizations (Young Academy)
- Funders and policymakers (representatives of VLAIO, Flemish government dep. EWI)
- AI practitioners

In a first phase, the Thinker-in-residence exchanged her vision and ideas on the Thinker's Cycle with the Cycle's initiators and the Steering Committee. The Steering

Committee had a first informal meeting on March 7, 2023, followed by a kick-off meeting with the Thinker-in-residence on March 8, 2023 at the KVAB. This kick-off meeting was organized for members of the KVAB, the Young Academy and other experts in the field to introduce Nowotny as Thinker-in-residence and the topic of the Thinker's Cycle. In this well-attended meeting, she gave a presentation based on her most recent book, entitled *In AI we trust. Power, illusion and control of predictive algorithms* (2021, Polity Press), followed by discussion and extensive debates.

Program KVAB meeting "AI as an Agent of Change," March 8, 2023, KVAB, Brussels

10:00-10:15:	Welcome & Introduction by Ine Van Hoyweghen, Chair Thinker's Cycle
10:15-11:00:	Lecture by Helga Nowotny "Liberal Democracies at Risk: Algorithmic Communication and the Delegitimation of Truth"
11:10-12:00:	Q&A with discussant Katleen Gabriels, Maastricht University

Based on these discussions, an outline of visions and ideas of the Thinker's Cycle was developed and discussed in a physical meeting by the Steering Committee. It was decided to focus on the most recent developments in AI, the impact of Generative AI on science and society. On the basis of this, the Thinker-in-residence prepared a first draft of the report.

In a second phase, the Thinker-in-residence was invited to participate in debates with Flemish experts, partners and stakeholders. She prepared relevant questions and discussed them with the Steering Committee in several online meetings during Spring 2023, where the further planning of the expert and stakeholder meetings was prepared for her upcoming September visits.

In order to put the Thinker in touch with leading Flemish experts in the field, an Expert Meeting was organized on Wednesday 13 September 2023 at the Academy. The experts were invited to read the draft of her report and were asked to present ideas and comments from their specific domains. She used the input of this meeting as feedback for her report and to develop concrete ideas for policy recommendations. The experts were also invited to write a reflection on the Thinker's report for this position paper (see section 3).

Program of the Expert Meeting "AI as an Agent of Change," September 13, 2023, Brussels

10:00-10:15	Welcome & Introduction by Ine Van Hoyweghen, Chair Thinker's cycle
10:15-11:00	Presentation of draft report by Helga Nowotny, Thinker
11:00-11:30	Presentation by Tinne Tuytelaers, KU Leuven

13:00-13:30 Presentation by Ann Dooms, VUB
13:30-14:00 Presentation by Johan Wagemans, KU Leuven
14:30-15:00 Presentation by Rosamunde Van Brakel, VUB/UHasselt
15:00-15:30 Closing debate and brainstorm for policy recommendations

Workshops with Flemish Stakeholders and Practitioners

The Thinker also entered into debate with various Flemish stakeholders, practitioners and partners. Although the discussions with stakeholders were flexible and open-ended, they were accompanied by a list of questions circulated ahead of each meeting. In addition, participants were invited to prepare written input in advance of the stakeholder workshops. As a result, the Thinker was gradually able to gain more insight into the local situation and could reflect on this from her international perspective. The input of these stakeholder workshops, marked by extensive note-taking, was used by the Thinker to develop concrete ideas for policy recommendations. These workshops were subsequently discussed in detail, and a summary of them is included in this Position Paper (see Section 5).

Stakeholder Workshop 1 on "ChatGPT and Teaching/Research," September 12, 2023, Brussels

Theme: AI as an Agent of Change: How are AI/ChatGPT used, experienced and supported in the formal education as well as in the information and training of the broader public in Flanders?

Stakeholder Workshop 2 on "AI Research and Applications," September 15, 2023, Brussels

Theme: AI as an Agent of Change: how are AI/ChatGPT used, experienced and supported in research as well as in the training of students in Flanders?

Taken together, these intense meetings, debates and workshops and their written input provided a realistic picture of the activities taking place in Flanders, in combination with the approaches, challenges, problems and prospects. From the perspective of her international experience, the Thinker then put together a basis of comparison for the Flemish context in order to develop policy recommendations. These policy recommendations were discussed with, and approved by, the Steering Committee on 21 September 2023.

Finally, the report and the recommendations of the Thinker-in-residence were presented at a well-attended public symposium 'AI as an Agent of Change' at the Palace of the Academies on Monday November 20, 2023.

During the symposium, an exhibition was organized as well on the impact of AI on Arts. Recent advancements in AI have strongly impacted the arts and the Steering

Committee (at the initiative of KVAB member Luc Steels) considered it interesting to highlight this facet with an exhibition featuring intriguing crossovers between AI and visual arts, poetry and music.

Program of Public Symposium, "AI as an Agent of Change," November 20, 2023, Brussels

9:00 REGISTRATION & EXHIBITION VISIT
9:45 Opening
Host/moderator: Jan Hautekiet
Welcome and opening words - Christoffel Waelkens, President KVAB
An introduction - Ine Van Hoyweghen, Chair of the Thinker's Cycle, KU Leuven
Presentation of the Thinker's report - Helga Nowotny, Thinker-in-residence
11:10 Impact of AI on Art
Panel discussion with artists followed by a Q&A with the audience
Chair: Luc Steels (in Dutch)
– Danny De Vos, visual artist
– Maarten Inghels, poet
– Andrew Claes, musician
– Kris Stroobants, conductor Frascati Symphonic
12:10 Insights of Researchers on Impact of AI
Presentations followed by a Q&A with the audience
– Ann Dooms, VUB
– Johan Wagemans, KU Leuven
– Walter Daelemans, UAntwerpen
13:30 How Is Flanders Doing in AI? - Bart De Moor, KU Leuven
The experience and outlook on future policy - Lucilla Sioli, Director for Artificial Intelligence and Digital Industry of DG CONNECT, European Commission

Exhibition curated by Luc Steels: Impact of AI on Arts

ARTISTIC CONTRIBUTIONS

AI & "La révolte des machines ou la pensée déchaînée"

Visual artist **Danny Devos** explores the use of AI in art in an exhibition of works generated by Artificial Intelligence Machine Learning Models. This has resulted in sculptural objects produced by 3D printing and CNC milling, involving electric motors and microcontrollers, based on the illustrations of **Frans Masereel** for "La Révolte des Machines."

Danny Devos (°1959), who lives in Antwerp, has presented his artistic endeavors throughout the world. Since 1979 he has done 160 performances in over 40 cities

in 12 different countries. For nearly forty years he has acted as performance, sound and "forensic" artist, purposefully making it his objective to remain a critical voice within and in opposition to the art scene.

Secrets

In 2021, artist **Luc Tuymans** and AI scientist **Luc Steels** collaborated to understand the process of art creation and art perception and interpretation. The result was featured in an exhibition at Bozar in April 2021. Here we show videos of Tuymans and Steels explaining these results and discussing their wider implications. This project was initiated by the EU Starts program, with the support of Gluon (Brussels) and Bozar.

Poem Booth

Maarten Inghels (former city poet of Antwerp) presents his "poem booth" during the symposium: it involves an experiment with generative AI, raising relevant issues in front of the audience. Language: Dutch.
Maarten Inghels made his debut in 2008 with Tumult in the Sandwich series, edited by author Gerrit Komrij, and has since developed into an original artist, poet and writer. His novel *Het mirakel van België*, about his experiences with the world's greatest master swindler, was published in 2021. His book *Contact* connected poetry, visual work and action. From 2016 to 2018, he was Antwerp's city poet.
A preview:

Kus elkaar, verliefden, onder de kruin,
Gegiechel galmt, vanuit de massa tuin.
Plots, een snorvogel schiet voorbij!
Herhaalt dit lied, dit zoete vrij.

AI Musicking

Andrew Claes and Frascati Symphonic provide a live performance of a new composition generated with AI and played by classical musicians.

Andrew Claes is a professional saxophone player and composer associated with the Royal Conservatory in Antwerp. One of his main projects AI Musicking is aimed at exploring innovative approaches to musical co-creation through machine learning. **Frascati Symphonic** is a collection of musicians from Leuven. They are well known for their performances of the classical repertoire, ranging from symphonic works and operas to chamber music. The orchestra is led by conductor Kris Stroobants. The musicians participating in a brief performance were Hrayr Karapetyan (violin), Delejan Breynaert (violin) and Shuya Tanaka (cello).

These various activities in 2023, including the symposium, made it possible for the public and other speakers to have several interesting interactions with the Thinker-in-residence. Based on these two-way communications, she has drafted her final report and the recommendations.

This Position Paper further comprises the report of the Thinker, the reflections from experts, the reactions from policymakers, the reports of the stakeholder workshops, and a closing chapter with conclusions and recommendations. The objective was to generate a broad set of recommendations that are relevant across disciplines and to contribute to future Flemish policy in the field. Drawing on the specific backgrounds, perspectives and competencies, then, this position paper presents no(t only a consistent analysis, but also valuable approaches and proposals for taking further steps. Given the quality of the Thinker's discussions with participants and her constructive findings, these efforts provide a solid foundation for the productive integration of AI into science and society in Flanders. It may even present insights, reflections, and recommendations that extend beyond Flanders, in particular to the SAM-SAPEA program for an accelerated uptake of AI in Science (SAM, 2023) and other actions in Europe and worldwide.

References

ALLEA (2023) *The European Code of Conduct for Research Integrity – Revised Edition 2023*. Berlin. DOI 10.26356/ECOC

Benjamin, R. (2019) *Race after technology: Abolitionist tools for the new Jim code*. Polity Press.

Bockting, C.L., van Dis, E.A.M., van Rooij, R., Zuidema, W. Bollen, J. (2023) Living guidelines for generative AI - why scientists must oversee its use. *Nature*. 622, 7984, 693-696. doi: 10.1038/d41586-023-03266-1

Crawford, K. (2021) *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, Conn.: Yale University Press.

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

European Commission (EC). (2021) Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. COM/2021/206 final. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

European Commission (EC) (2023) 2023 State of the Union Address by President von der Leyen Strasbourg, 13 September 2023, State of the Union Address by President von der Leyen (europa.eu)

Federal Public Service, Policy and Support (BOSA), (2022) National Convergence Plan for the development of Artificial Intelligence (2022) Plan AI (NL)-compressed.pdf (belgium.be)

Ferrari, F., van Dijck, J. and van den Bosch, A. (2023) Foundation models and the privatization of public knowledge. Nat Mach Intell 5, 818-820. https://doi.org/10.1038/s42256-023-00695-5

Human-Centred Artificial Intelligence (HAI) Stanford University, Home | Stanford HAI

Jones, N. (2023) How to stop AI deepfakes from sinking society - and science. *Nature.* 621, 7980, 676-679. doi: 10.1038/d41586-023-02990-y

Kalluri, P. (2020) Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*. 583, 7815, 169. doi: 10.1038/d41586-020-02003-2

Lazar, S. and Nelson, A. (2023) AI safety on whose terms? *Science*. 14, 381, 6654, 138. doi: 10.1126/science.adi8982.

Nature Editorial (2023) Stop talking about tomorrow's AI doomsday when AI poses risks today, *Nature* 618, 885-886. doi: https://doi.org/10.1038/d41586-023-02094-7

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

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019), Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 25, 366, 6464, 447-453. doi: 10.1126/science.aax2342.

OECD (2023) Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research, OECD Publishing, Paris, https://doi.org/10.1787/a8d820bd-en

Rabaey, J., van Est, R., Verbeek, P.P., and Vandewalle, J. (2020) *Societal values in digital innovation: who, what and how?* - KVAB Thinker's Programme 2019, KVAB Position paper 66 b.

Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 1, 206-215. https://doi.org/10.1038/s42256-019-0048-x

Scientific Advice Mechanism to the European Commission (SAM). (2023) Artificial intelligence in science. Scoping paper, 4 July 2023. https://scientificadvice.eu/advice/artificial-intelligence-in-science/

Steels, L. et al. (2017) *Artificiële intelligentie. Naar een vierde industriële revolutie?*, KVAB Standpunt 53.

VAIA (2019) Flemish Policy Plan for Artificial Intelligence. Flemish Policy Plan AI - VAIA - Flanders AI Academy

Vandewalle, J., Acheroy, M. et al. (2022) A call for an accelerated digital transformation for Belgium, ARB/KVAB Position paper 77.

van Dis, E.A.M., Bollen, J., Zuidema, W., van Rooij, R., and Bockting, C.L. (2023) ChatGPT: five priorities for research. *Nature*. 614, 7947, 224-226. doi: 10.1038/d41586-023-00288-7.

Van Noorden, R., Perkel, J.M. (2023) AI and science: what 1,600 researchers think. *Nature*. 621, 7980, 672-675. doi: 10.1038/d41586-023-02980-0

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

2. Report of the Thinker

AI as an Agent of Change

Helga Nowotny, Thinker-in-residence

Elizabeth Eisenstein's influential classic *The Printing Press as an Agent of 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 behavior 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 the 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 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 toward more standardization and greater efficiency, but also toward 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 the 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 played by the down-sized small electric motor 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-away 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 US and China. Among other things, 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, like 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 Thinker's 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 agency 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 toward 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, however, these risks pale compared to those AI-powered battleships without crews or self-directed drone swarms, which 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. 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, and 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 backward 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 leeway that exists 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, and 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 analog (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 DIKW pyramid 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 upward 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 "deficit 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 among themselves the educational efforts directed toward 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/

Toward 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 interdisciplinarity, 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 Thinker's 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), 321-337.

Uzogara, E. (2023) Democracy Intercepted. Did platform feeds sow the seeds of deep divisions during the 2020 US presidential election? *Science* 381, no. 6656, 28 July, 386-387.

3. Reflections from experts

Large Language Models: The Rise of the Daydreaming Zombies

Walter Daelemans, University of Antwerp

Helga Nowotny's text "AI as an Agent of Change" provides a welcome reminder of the unpredictable consequences a new technology can have on society. The best we can do is discuss scenarios for unintended consequences and prepare to adapt to them. But doing this for AI is not without problems because it is not a new technology, and its label is used for a diverse range of applications. I will argue that to alleviate negative consequences we should invest more in public AI research rather than trying to halt it.

We already have many examples of good and bad AI. We have autonomous rovers and helicopters on Mars, but we also have the first autonomous weapons in use, and the UN not yet succeeding in having them banned. We have deep fakes, but we also have more than a decade of experience with usable Machine Translation. This has not adversely affected the employment of translators; instead, it has opened up new markets and applications. These systems provide an enormous boost for communication and understanding across language barriers, from international trade to science and entertainment, as well as to aiding refugees and migrants. Speech synthesis (text to speech) and speech recognition (transcription) have also been a significant help for people with perceptual disabilities and will have mainly positive effects in many new contexts thanks to further quality improvement.

The sudden introduction of ChatGPT to the general public in late 2022 caused significant concern about the dangers of AI, so I will focus here on those Large Language Models (LLMs), and more generally Generative AI (GenAI), as well as on scenarios in which they play a role, either for good or bad.

Extinction by AI Is Not Imminent

LLMs are based on a long research tradition in Natural Language Processing (NLP). Statistical language models, the ancestors of ChatGPT, have been responsible for the first usable machine translation and speech recognition applications from the 1990s onward. A statistical language model assigns a probability to a sequence of words. This is useful when the best translation or the best speech transcription must be selected among many possibilities. However, exponential scale increases, both in the size of the language model (number of parameters) and in terms of the amount of data on which they are trained, have made a huge difference. LLMs now show "emergent" behavior that reaches far beyond assigning probabilities to chunks of text. The models understand, generate, translate and transform long texts, change their style or complexity, summarize them and answer questions

about them. They are good at programming, can be empathetic and creative, and show commonsense reasoning. They can be said to "understand" text and to be "intelligent," even though that is of course mainly a (philosophical) terminological issue. If understanding and intelligence are functional concepts (like flying), their simulation is for all purposes equivalent to the real thing. Just like an airplane achieves the function of flying, differently from a bird, LLMs can functionally "understand" text without being human. But LLMs see the world only through the filter of text written by people, both fiction and non-fiction. They clearly do not have agency, emotions, consciousness, opinions or goals (apart from the goal of providing an optimal completion to a prompt), even when they say they do, and they live in a dream world without grounding in reality (hence "daydreaming zombies," close relatives of the philosophical zombies). It is also hard to see how grounding and self-awareness could emerge from the objective functions with which language models are currently trained, even with more or multimodal data. If the "extinction of humanity" threat by superhuman intelligence seems far away still, there are more urgent worries.

Disappearing Skills

The way in which LLM-based AI will become prominent in society is in the form of assistants. Teaching assistants, study assistants, programmer's assistants, doctor's assistants, etc. In the mid-term, this will not necessarily have adverse effects on employment or on our view of what is worthwhile to be studied. As is the case in Machine Translation, productivity and quality will increase, paving the way to new opportunities. AI may even have a democratizing effect. If you are an expert programmer or copywriter, you will run the risk of losing that status because everyone in your trade will be able to level up with high quality products. But in the long term, with even more advanced AI models, society may have to adapt to unseen levels of automation of cognitive tasks and a reluctance to go through the trouble of gaining expertise in skills AI systems are better at anyway. It is hard to imagine now, but human writing and programming may become as obsolete as mental arithmetic.

The Need for Advanced Spam Filtering

Of immediate concern is the way LLMs can be used to influence, persuade, attack and manipulate people by combining text- and image-based GenAI. These are fake humans created by real humans. Society will have to decide whether to prohibit this, and if so, how. In any case we will need AI research, by the public sector, to develop systems that can detect and fight unwanted AI-generated content. This is not an impossible task. Investment is also needed in developing better approaches to validation, quality control, guardrail development, and explanatory capabilities applied to LLMs. To achieve this, it is necessary that the AI research community has access to open source LLMs that are comparable in size and capabilities to the commercial models. There is a role for Europe to make this happen.

Before ChatGPT, many viewed the goals of AI with skepticism or considered them distant promises. Now that some of these goals have been achieved, be it in unexpected ways, we should not abandon the research just because of fear of abusive uses. LLMs and GenAI do not tell the full story and arguably they are not as powerful as some think they are. Yet, they are interesting enough to be researched, analyzed and experimented with. Only then will we be able to build in safeguards and fight abuse.

A Behavioral Science Perspective on AI

Jan De Houwer, Ghent University

Whereas behavioral science is typically concerned with the behavior of individual organisms such as a human or other animal, one could look also at other systems from a behavioral science perspective. For instance, one could argue that an AI (such as a computer algorithm or artificial neural net) is a system that behaves: it changes its state (e.g., its output and/or the strength of the links in its network) as the result of events in its environment (e.g., the input it receives from users; see De Houwer & Hughes, 2023). Based on the premise that AI is a behavioral system, one can deploy methods, concepts and insights from behavioral science to the study of AI (Rahwan et al., 2019). This allows us to examine in a systematic way the similarities and differences between the behavior of AIs and other systems even when the mechanisms underlying the behavior of the different systems is fundamentally different or as yet unknown (as is the case, for instance, in deep learning models).

A behavioral perspective on AI highlights the many parallels between AI and other behavioral systems such as individual organisms. Like individual organisms, AIs cannot only respond to events in their environment but, based on experience, also change the way they respond to an event (i.e., AIs can learn; cf. De Houwer & Hughes, 2023). Via their behavior, AIs can also change the environment of other systems in such a way that those other systems also change their behavior (e.g., when an AI presents text in response to a query of a person, this can change the behavior of that person). Moreover, AIs can do so in an individualized way (i.e., based on information about the other system) and instantaneously (i.e., responding online to the behavior of the other system). Guided by known behavioral principles (e.g., reinforcement; see Catania, 2013), AIs thus can be programmed or trained as a tool for behavior change. These powers stand in marked contrast to those of older technologies such as the printing press. Such older technologies offer a way to change the behavior of people by changing their environment (e.g., creating the opportunity for people to read books) but lack the capacity to respond to the environment, to change the way they respond to their environment, and thus to dynamically influence the environment and behavior of

other systems. From a behavioral perspective, it makes more sense to compare AIs not with a technology such as the printing press but with a behavioral system that uses technology. Before the advent of AI, technologies were used by systems comprising of a single person or a group of persons. For instance, the printing press was and is used by individual or groups of novelists, philosophers, scientists, marketeers and so on with the aim of influencing the behavior of people in certain ways (e.g., adopting ideas, buying products). As a behavioral system, AIs can use technologies much like a system composed of humans would (e.g., generate text, decide on who to expose to which text, reinforce people to behave in certain ways). AI remains a technology in that it is designed and maintained by humans to perform certain tasks, but even the design and maintenance of AIs can, at least in principle, be performed by AIs. In the latter case, AIs would qualify as autonomous behavioral systems, much like individual organisms would.

A behavioral perspective on AI does not reveal only similarities between AIs and individual organisms, however. A fine-grained analysis of the behavior of current AIs reveals crucial differences with the behavior of humans. One crucial difference lies in the capacity to behave symbolically. Humans have a unique capacity in responding to one thing as-if it is related in a certain way to something else. For instance, they can act as-if the word "GLASS" refers to a physical glass even though the relation between both is arbitrary and defined by social convention or act as-if a dime coin is more than a nickel coin in terms of monetary value even though a nickel is more than a dime in terms of physical size. Many scientists have argued that this type of symbolic behavior lies at the core of human cognition (see McLoughlin et al., 2020, for a review). Some have also mapped out (in broad strokes) the extensive learning history that humans have to go through before they can display this type of behavior (e.g., Hayes et al., 2001). There are good reasons for saying that current AI systems do not show symbolic behavior. ChatGPT, for instance, cannot give a sensible answer to the following question: "Assume that yellow is more than blue and that red is less than blue. Is red more than yellow?" A verbally-able human can tell you that red cannot be more than yellow because he or she can act as-if yellow is more than blue and red is less than blue. ChatGPT has never received the training that is necessary to show symbolic behavior. It has been fed masses of data and trained to construct responses that sound reasonable on the basis of this data but it has not reached the symbolic stage. Of course, future AIs might be able to show symbolic behavior if they are trained in a way that is similar to the learning history that produces symbolic behavior in humans. As such, a behavioral perspective on AI cannot only shed light on the current nature of AI but will also help shape the future of AI.

References

Catania, A. C. (2013). *Learning*. Sloan.

De Houwer, J., & Hughes, S. (2023). Learning in individual organisms, genes, machines, and groups: A new way of defining and relating learning in different systems. Perspectives on Psychological Science, 18, 649-663. https://doi.org/10.1177/17456916221114886

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (Eds.). (2001*). Relational Frame Theory: A Post-Skinnerian account of human language and cognition*. New York, NY: Plenum Press.

McLoughlin, S., Tyndall, I., & Pereira, A. (2020). Convergence of multiple fields on a relational reasoning approach to cognition. *Intelligence*, 83, 101491. https://doi.org/10.1016/j.intell.2020.101491

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A., . . . Wellman, M. (2019). Machine behaviour. *Nature*, 568, 477–486. https://doi.org/10.1038/s41586-019-1138-y

Who Will Be the Guardian Angel in the Footsteps of Erasmus?

Marc De Mey, Ghent University

The book *The Printing Press as an Agent of Change*, Eisenstein's (1979) treatment of the fifteenth-century printing revolution, is interesting because it differentiates between impacts of the new technology on culture, religion and science. The impact of the introduction of the printing press in society is not to be reduced to a simple effect of scale. Science is differently affected compared to religion and literature.

The Nowotny-report refers to science with the Eisenstein quote: "The fact that identical images, maps, and diagrams could be viewed simultaneously by scattered readers constituted a kind of communication revolution itself" (Eisenstein, 1980, p. 53). "A communication revolution in itself," the quote might seem to mention it cursory, but Eisenstein means a genuine qualitative change affecting the accretive nature of scientific progress. Published scientific books of a single edition, through being surely identical, assure any participant entering the field an identical representation of the state of affairs in that field. Putting every participant on an equalized ground ("uniform grid," Eisenstein, 1980, p. 517) eventually allows the single potential innovator to surpass the attained level and take the field one step further. The equalizing effect of printed editions, instead of fluctuating and unstable scribe products, is essential for the "*cumulative cognitive advance* and incremental change" (Eisenstein 1980, p. 412, italics added) that characterizes genuine scientific growth.

By pointing out in the report, just after the Eisenstein quotation, that "(i)t resulted in a veritable knowledge explosion in the 16th century" emphasizes especially scale without specifying the mechanism of content stabilization at disciplinary level. In the next paragraph: "It is therefore tempting to draw parallels between the knowledge explosion of the 16th century and the 'information explosion' that holds us in a firm grip since some time." The sentence that follows has equally the emphasis on scale: "(t)he recent public release of Generative AI based on Large Language Models *has merely added* to the overwhelming abundance of possibilities that AI has opened" (italics added). It might be premature to speculate already on the differential qualitative impacts AI might have on science, literature, art and culture in general. Nevertheless, if we envisage specific measures to optimize its potential contributions, it is indicated that those measures are adapted to the specific nature of the domains involved. Do we witness AI having differential impacts comparable to the ones Eisenstein documents for religion and science?

Take first religion. As Leitmotiv for the fourth chapter, Eisenstein invokes a passage from A.G. Dickens' *Reformation and Society in Sixteenth Century Europe*, pointing out the important role of the printing press: "Between 1517 and 1520, Luther's thirty publications probably sold well over 300,000 copies …Altogether in relation to the spread of religious ideas it seems difficult to exaggerate the significance of the Press, without which a revolution of this magnitude could scarcely have been consummated. Unlike the Wycliffite and Waldensian heresies, Lutheranism was from the first the child of the printed book, and through this vehicle Luther was able to make exact, standardized and ineradicable impressions on the mind of Europe. For the first time in human history a great reading public judged the validity of revolutionary ideas through a mass-medium which used the vernacular languages together with the arts of the journalist and cartoonist" (quoted from Eisenstein, 1980, p. 303).

In the pages that follow, Eisenstein refers to data indicating that Luther's ninety-five theses translated in German and printed in substantial numbers circulated in Nuremberg, Leipzig and Basel already by the end of 1517. This is in only three months after Wittenberg! The printers were apparently quite eager to produce and sell whatever short piece of text they assumed to have commercial appeal. Eisenstein quotes Zwingly who recommends, in 1519, the Luther tactic: offer in door-to-door sale just one single article so that the potential buyer is saved from choice conflict and simply has to decide "yes, I buy" or "no, thank you." This is not unlike the business zeal with which ChatGPT is now pushed down the throat of pc- and internet-users worldwide: "Ask me any question." Luther sensed the potential of the printing press and in 1522 he promptly came with a translation of the *New Testament* in common language and in a large edition that rapidly sold out. Codi Byte publishes in 2023 a *ChatGPT Bible* with as subtitle "Everything You Need to Know about AI and Its Applications to Improve Your Life,

Boost Productivity, Earn Money, Advance Your Career, and Develop New Skills." It has ChatGPT for creativity, for entrepreneurs, for researchers, for educators, for writers, for programmers, for professionals, for social media managers, for journalists and for linguists. Presented as such, ChatGPT is a brazen commercial endeavor meant to invade the daily life of internet users over a broad range of activities, from finding a recipe for cooking eggs to writing fiction or debugging computer code. In some browsers this ChatGPT has introduced itself to the user as "your co-pilot," capable of helping you in all what you undertake, your genuine digital twin. Is there another ChatGPT?

Eisenstein draws a comparison between the historical figures of Luther and Erasmus. Both set out to remedy dysfunctional situations in the Church and both saw the potential of print as proper instrument. However, where Luther used it to re-present his religious message in generally understandable form to the public at large, Erasmus saw it has an opportunity to improve the quality by deepening the scientific study of the texts to be enshrined in the new medium. Given the opportunity to spread God's word in a superior material form (print), it had to be done in the appropriate way with serious scholarly study of the original languages involved. Therefore, he inspired and supported the founding of the *Collegium Trilingue* (1517!) at Louvain University for the study of Latin, Greek and Hebrew. Currently, there are comparable concerns about the premature spreading of the products of AI such as ChatGPT, while there are similar initiatives like *International Institutes for Advanced AI*, co-authored by, among several others, one of the pioneers of deep learning Joshua Bengio, for in-depth study of AI, including consciousness https://arxiv.org/pdf/2307.04699v1.pdf. The recommendations of Helga Nowotny specify our local prerequisites for conjoining in such global actions.

AI as an Agent of Change – Seen through the Eyes of a Mathematician

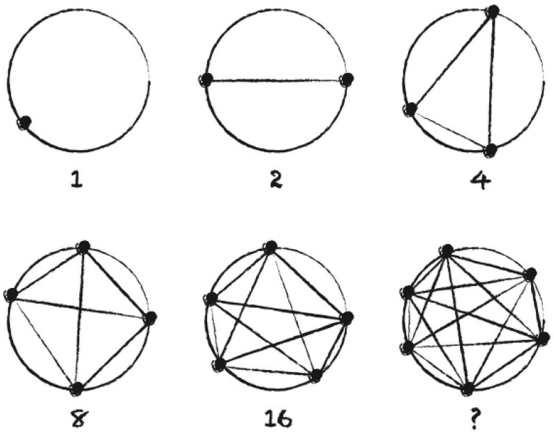
Ann Dooms, VUB

To understand where AI is at today and how it could evolve in the future, we need to traverse the corridors of time.

Going back 5000 years we encounter the Babylonians that were driven by an innate curiosity to understand the world through numbers. From their clay tablets, we know that they gathered massive amounts of data from meticulous and prolonged observations. Though lacking the mathematical formalisms of later eras, they laid the groundwork for recognizing numerical patterns in data from which they could make accurate predictions, ranging from astronomical phenomena up to properties of mathematical objects. They already grasped the notion of right-angled triangles and the relationship between the lengths of their sides, now referred to as the Pythagorean Theorem. On the Plimpton 322 tablet they assembled a list of so-

called Pythagorean triples, which are integer solutions of the equality $x^2+y^2=z^2$. Did we wrongly attribute the famous theorem?

To answer that question, I will let you delve into pattern recognition. Choose two points on the border of a circle and connect them with a line segment. This divides the circle into two regions. Now, by choosing a third point and connecting it to the first two, you will get four regions. Each time you choose a new point and connect it to all previous ones, you get additional regions. In the drawings below, you can see that the number depends on the positioning of the points on the border. Moser's Circle Problem from 1949 questions how many regions you can maximally obtain for a given number of chosen points. At first sight, we observe a pattern that relates the number of points with a power of 2 regions. Can we use this to predict the number of regions for 6 points?



Number of points	Number of regions
1	$1 = 2^0$
2	$2 = 2^1$
3	$4 = 2^2$
4	$8 = 2^3$
5	$16 = 2^4$
6	$32 = 2^5$?

Although it is tempting to think the problem can be captured by this nice numerical relationship, it is unfortunately wrong as one can prove that for 6 points there are maximally 31 regions. This example clearly shows the danger of generalization. The mere presence of an apparent pattern does not necessarily imply it holds. The Babylonians observed the relationship between the lengths of the sides in a right-angled triangle, but Pythagoras and his pupils were the first to formulate proof of the universal correctness, hereby giving birth to mathematics as a science in its own right. The only science in which you can obtain uncontested truths by giving logical proofs of statements.

The Greeks set the scene for the advancement of mathematics, crafting it into a powerful toolkit with instruments that forever hold their value. In the 17th century, the cosmos surfaced as the canvas for new mathematics based on the observations of the Babylonians. Johannes Kepler's laws of planetary motion

and Galileo Galilei's telescopic observations shattered the geocentric worldview, unveiling a celestial ballet governed by universal laws beautifully captured in geometrical formulas. Enter Isaac Newton, whose laws of motion and universal gravitation provided an explanation for the observed dynamics. To formulate them he made a leap in abstraction by, together with Gottfried Wilhelm Leibniz, introducing the mathematical concept of derivative as a tool to study motion. As such, the *derivative* literally became an *agent to measure change*.

This jumpstarted, in its turn, a branch of mathematics called calculus, which deals with continuous change to help understanding and analyzing varying quantities. It aids in particular to approximate functions with prototypes that are easier to understand and calculate with. This made it possible, for example, to create tables of sines and cosines up to a desired precision that not only facilitated practical applications like navigation. To perform the easy but tedious calculations Leonardo Da Vinci laid the foundation for a mathematical machine: the mechanical calculator, later commercialized by Blaise Pascal and operated by humans called *computers*.

As the clock ticks into the 19th century, we meet the astronomer Charles Babbage who was annoyed by the spurious but dangerous errors occurring in the tables of approximated functions. Human mistakes in operating the calculators or writing down the results could lead to drastically wrong computed navigation routes. The Industrial Revolution, driven by steam, led him to invent a calculator, the *Difference Engine*, that performed computations autonomously. To allow more advanced, in particular, combined calculations eliminating the human in the loop, he teamed up with Ada Lovelace, Lord Byron's daughter. Together they conceptualized the first programmable machine, inspired by the steam-powered Jacquard loom that could produce fabulously looking complicated patterns by making use of punch cards to steer the machine. Unfortunately, they never saw, what they called the *Analytical Engine*, into action.

It was only when the torch was passed to the visionary Alan Turing a century later that the digital age was truly born. In 1938 he mathematically proved that one could devise a machine that can compute everything a human can do by hand, a theoretical construct that sparked the development of the electronic computer as we know today. After cryptographic endeavors during World War II, he further shaped the history of computing in 1950 with his seminal paper *Computing Machinery and Intelligence*. With the *Imitation Game*, now known as the *Turing Test*, he created a framework for artificial intelligence, in which we are trying to make machines that can learn to solve problems the way we humans do. From trial and error up to learning by example.

Turing's mathematical principles gave rise to the now hugely popular neural networks, inspired by the intricate web of biological neurons in our brain. An artificial neuron will output information when it gets enough stimulus from his

input. Frank Rosenblatt proved in 1958 that it can learn when to "fire output" from seeing examples in which it should. To solve more complex problems, improvements were steadily made by combining artificial neurons into networks, culminating in a mathematical feat in 1975 that proves that this model can still learn from examples. The key is Newton's derivative to control movement in the parameters of the network during training. The so-called *backpropagation* algorithm corrects intermediate inaccurate predictions by traversing back through the network and adjusting numbers in the right places. Although it was only due to the massive uprise of digital data and computational power that one could implement such neural networks to discern complicated patterns in data far beyond what a human could compute by hand in a lifetime and leading to mesmerizing generative tools such DALL-E and ChatGPT.

But … we are still quite distant from machines that can learn and think the way humans do. One can easily fool the current products, exposing them as probabilistic parrots that cannot reason over the learned content. Will they ever be able to prove that predicted mathematical patterns always hold? In this quest we will once again witness the fascinating interplay between mathematics and technology in uncharted territories. AI will definitely be an agent of change in mathematics and the other way around.

Should There Be a Right to Refuse?

Katleen Gabriels, Maastricht University

On 20 October 2023, Belgian newspapers reported that, for the first time, grid operator Fluvius will take a refuser of a smart electricity meter to court. It is delicate to start from an example for which I do not know the underlying facts, but this case raises compelling questions about how much agency and room for refusal users still have in a society saturated with digitalization and AI. If people have a valid reason not to use a "smart" technology, for example because it is at odds with their privacy, to what extent do they have a right to refuse?

Users are of course neither powerless nor passive. In 2011, Maximilian Schrems requested all the data that Facebook kept about him. Facebook was, due to European privacy legislation, obliged to provide these to him; every European citizen has the right to access data collected about them. Facebook turned out to keep more than 1200 pages about Schrems. To protect the privacy of citizens, American companies are not allowed to transfer personal data of Europeans for commercial purposes without their consent. The positive side to this story is that one person can successfully reveal the practices of a large and powerful global player. Schrems became a lawyer and privacy activist.

Next to activism, there are various ways to resist and refuse technology and it does not necessarily mean that the technology is not used at all. Brunton and Nissenbaum (2015) encourage data obfuscation, to give users more agency to protect their privacy online. In doing so, people can still use the technology, but their data are protected better, for instance, by encrypting them. Users also often "tweak" or "fit" technologies to their own standards. Kamphof (2015) observed how professional caregivers and developers actively try to fit smart monitoring technologies into patients' daily practices and implement strategies to respect their patients' privacy. Developers tend to focus more on safety and autonomy enhancement and less on physical privacy (Birchley et al., 2020).

Earlier in 2023, the Italian start-up company Cap_able has launched its Manifesto collection: the knitted garments have been deliberately designed to confuse and trick facial recognition software, such as cameras on the street. For instance, instead of recognizing a person, the camera "sees" an animal that is embedded into the pattern. In doing so, the company gives the wearer more options, namely, to be anonymous. Cap_able seeks "to be an exemplary leader in raising awareness of the importance of one's rights: a means to express oneself, one's identity and the values shared within a reference community." They also want to increase awareness of "misuse of facial recognition technology."

Companies such as OpenAI do not respect the copyright of all the texts and images with which they train their Large Language Models, including ChatGPT. Nightshade is a recent "data poisoning tool" to give more agency to artists: it adds noise to their digital art to prevent big companies to use it to train generative AI technologies such as Midjourney and Stable Diffusion.

Resistance to technology is a multifaceted concept, encompassing various levels of engagement. People may exhibit resistance in several ways: refusing, rejecting, data obfuscation or 'poisoning', fitting, tweaking, cheating (e.g., data of activity tracker), negotiating/lobbying and protesting, or simply by expressing concerns. Public discourse and public space play a pivotal role in shaping AI technology. When users express their concerns and resistance to AI, it sparks important conversations about the ethical implications of AI adoption. These discussions have already led to regulatory changes (e.g., in the case of Schrems), increased transparency and improved accountability in the tech industry. Users, therefore, can act as catalysts for societal reflection and change.

In an age where AI technology is deeply intertwined with our daily lives and public infrastructures, the power to refuse AI is a fundamental expression of user agency and autonomy. For that reason, the right to refuse needs more attention. This right already exists in different forms, for instance, to refuse unsafe work, or the right to strike, which is also a form of refusal and resistance. Yet, the right to refuse certain forms of AI technologies in public infrastructures might play a

role in ensuring that AI technology respects users' autonomy, privacy and ethical considerations, ultimately guiding the future of AI in society. How can such a right look like? And which form should it take?

References

Birchley, G., Huxtable, Murtagh, M., ter Meulen, R., Flach, P., & Gooberman-Hill, R. (2020). Smart Homes, Private Homes? An Empirical Study of Technology Researchers' Perceptions of Ethical Issues in Developing Smart-Home Health Technologies. *BMC Medical Ethics* 18(23).

Brunton, F. & Nissenbaum, H. (2015). *Obfuscation: A User's Guide for Privacy and Protest*. MIT Press.

Kamphof, I. (2015). A Modest Art: Securing Privacy in Technologically Mediated Homecare. *Foundations of Science* 22, 411-419.

The Borg Society

Yves Moreau, University of Leuven

> *We are the Borg. Lower your shields and surrender your ships. We will add your biological and technological distinctiveness to our own. Your culture will adapt to service us. Resistance is futile.*
>
> *The Borg. Star Trek: First Contact*

To explore how AI acts as an agent of change in human society and to what extent it will empower human autonomy or erode it, let us examine its interactions with society. AI systems can be thought of as "cognitive machines" that process text, speech and images on a large scale in a human-like fashion. They interact with humans and thus inevitably influence them.

To lay the foundation for our discussion, let us begin with some fundamental concepts. A (mechanical) machine is an engineered system of parts that uses power to transmit forces, motion and energy in such a way as to produce a predictable and desired output in a manner determined by a specific input. The first key aspect of a machine is the modulation of forces and motion to achieve a desired effect in the physical world. Leaving the biological realm and "molecular machines" aside, the second key element of a machine is that it is engineered—i.e., designed and built to achieve a specific purpose. The third aspect is that most machines are tools, which means that they perform some productive task (by contrast, Rube Goldberg machines and race cars are machines but not tools).

Machines amplify human capacities, in particular in repetitive labor. Machines were the beating heart of the Industrial Revolution.

Similarly, computers are "information machines." The key difference is that instead of mediating forces and motion, they process information (in the form of electromagnetic signals). They do so through algorithms, which are sequences of steps to achieve a desired output, in ways that are not too dissimilar from what mechanical machines do. Early computers, from Charles Babbage and Ada Lovelace's Difference and Analytical Engines to Konrad Zuse's Z3, were actually (electro)mechanical machines. While the advent of the transistor radically changed the design and relevance of computers, the analogy with mechanical machines is such that by now digital computers are called machines too. Computers are central to our Information Revolution. I explored the parallel between the printing press as a key driver of the Protestant Reformation and the internet as the driver of the Information Revolution in the essay "The Geek Reformation."[1] This parallel is a useful way to identify plausible patterns in the current state of affairs – even if the limitations of historical analogies must be acknowledged.

With the term "cognitive machine," or "cog" for short, we want to emphasize some key characteristics of large-scale AI systems: they process large amounts of data (Big Data), in particular unstructured and poorly structured data, such as natural language, speech and images, and interact with users through such modalities, which creates the perception of human-like capabilities. Moreover, they are highly scalable and processing more data only requires adding more computing blades or firing more cloud servers, which provides favorable economies of scale. Importantly, the number of people needed to run such systems grows much more slowly than the amount of data processed or the number of customers. Nevertheless, cogs are in a sense "just" computers. While we could try to delineate the difference between AI systems and "classical" computing, our focus is on how large-scale computation affects society.[2] Whether or not a particular cog meets the threshold to be considered "artificial intelligence" is not the core issue.

It is the extraordinary scalability of cogs that enables the "connectives," which are the structures that emerge from the networking of cogs with their users, providers and designers. Cogs often interact with, and thus influence, a large number of users. Because they exist in a competitive economic environment, two of their key features are that they are designed to (1) leverage network effects to acquire and retain users and customers or (2) compete for user attention and

---

[1] Yves Moreau (2016) The Geek Reformation, in *A Truly Golden Handbook: The Scholarly Quest for Utopia*, Leuven University Press, 464-477.

[2] Kate Crawford (2021) *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*, Yale University Press.

impact user behavior – so as to maximize reach and revenue. Social networks, such as Twitter/X, Facebook or TikTok, are connectives that leverage both of these aspects. Market platforms, such as Amazon or Uber, mostly leverage network effects between providers and customers. Google Search leverages network effects among advertisers to generate the revenues needed to provide a service that outcompetes those of its rivals. OpenAI aims to become the engine that powers AI text generation across a myriad of businesses. A key aspect of connectives is that they allow for extremely compact organizations with immense reach. OpenAI reached hundred million active users within two months of publicly launching ChatGPT at the end of 2022 with only three hundred or so employees. It has recently been valued at over $80 billion. In the connective, the cog, top management, and a few engineers form the brain of a giant octopus whose tentacles reach hundreds of millions. As such connectives get further boosted by increasingly sophisticated AI that decrypts, anticipates and creates every human need and desire, expect them to further insert themselves into every nook and cranny of the human experience. Although this metaphor appropriately suggests immense power for those who control the brain of the octopus, feedback mechanisms from users exist in the form of market forces, which seriously constrain the options of those making decisions.

If connectives deploy themselves ever deeper into society, what kind of society can we expect to see emerge? What agency will AI have and what will its impact be on human agency? If we define agency as the ability to set goals, sense one's environment, and plan and carry out a reasonable course of action to move toward those goals, one can argue that many engineered systems – from the humble thermostat to the AI managing a stock market portfolio – show some degree of agency (if we accept for a moment that agency does not require consciousness). At this point, it is appropriate to bring another metaphor forward to describe the relation between humans and their machine and tools: the cyborg, which is an "organism that has restored function or enhanced abilities due to the integration of some artificial component or technology that relies on feedback."[3] The notion of feedback is important here. Someone with a pegleg would not qualify as a cyborg, while someone with an advanced myoelectric prosthesis technically would. Typically, the modern archetype would involve direct connection and feedback between the prosthesis and the brain, something already found in cochlear implants. Furthermore, in modern cyborgs, cognitive functions are often in part carried out or enhanced by computer systems. The key idea is that in a cyborg, the locus of agency can be situated in either or both the biological and digital parts of the brain. We are not considering here a potential cyborg future for human beings, but rather what happens when the loci of decision-making across society become split between the human and the digital.

---

[3] Joseph Carvalko (2012). *The Techno-human Shell-A Jump in the Evolutionary Gap*, Sunbury Press.

To consider how connectives affect society, we invoke the relentless antagonist from Star Trek: the Borg, which is a collective of cyborgs united via a "hive mind" (also called "collective consciousness") and led by the Borg Queen. Although the queen leads the collective and is the only member with personal autonomy, she is also the product of all experiences and memories of all members of the collective.

In a society where all individuals are constantly connected to cogs and each other via connectives, where is the locus of decision-making? If an AI suggests a great gift for your significant other's birthday by assessing their digital trail, did you really make that decision just because you approved a suggestion far better than any idea you could have come up with? When an AI system screens CVs for a job interview, is the decision-making locus not already shared between human and machine? If an autonomous AI weapon selects and kills an enemy soldier, was the locus of decision-making really with the operator who provided a broad objective ("take this military compound") or with the developer who programmed the AI decision-making framework without any knowledge of the specifics of this particular decision (and often no understanding of how specific decisions are being made)? We can argue that to a certain degree this process of "displacement of agency" toward digital systems is already at play in our societies. How often are we not confronted with the obtuse "Computer Says No" in our daily lives, where we could probably have wiggled our way through half a century ago but now face an inscrutable computer system or a human bound by it? Digital systems already shape significant aspects of the architecture of modern society. They are simply so boringly built into the fabric of society that we barely pay attention to them – except for the occasional curse at a screen. While social relations, norms, markets, legal systems, and bureaucracies have shaped societies from ancient history to late modern history, these structures were solely embedded in human agency. With the advent of the mainframe, the personal computer, and then the cloud, we have started to see a shift of agency from the human to the digital realm. How many people have lost their job because of an automated formula applied by some management consultant – or even a bug in an Excel sheet? How many couples have been formed or dissolved because of algorithmic suggestions? Stock markets have collapsed under algorithmic trading without anyone really understanding why.[4] These trends will be boosted exponentially and in countless ways by the emergence of ever more powerful AI cogs. I put forward the thesis that we are moving towards a Borg Society, where the enmeshment of human and digital agency will be ever tighter and impossible to disentangle. Just as it has been historically nearly impossible to disentangle individual agency from resultant socio-historical forces, it will become impossible to decide where agency is human or digital.

[4] https://en.wikipedia.org/wiki/Black_Monday_(1987)

Although doom scenarios are not the most helpful and current predictions of a Skynet/Terminator AI existential risk might be a distraction from more credible concerns, we can reasonably worry about how far this displacement of agency will go. Human autonomy is already substantially constrained by social structures. Even where we are free to make decisions, which decisions we can make and what are the available options is greatly constrained. How much further would this autonomy shrivel in the Borg Society? Who will see the amazing possibilities offered by AI enhance their individual autonomy and who will become a tool of digital systems, carrying out whatever substandard non-automated tasks remain? Who will be left on the side of the road and rendered utterly obsolete? Are Silicon Valley billionaires our Borg Queen? Will we be fully assimilated within a few generations into a society that is totally unrecognizable to the present us? Are humans even needed in the long run for the Borg Society's continued existence? Can these historical dynamics be mitigated or shall we accept that resistance is futile?

Turing's Curse

Luc Steels, VUB AI Lab Brussels

The Turing Test

In a famous paper published in 1950, Alan Turing proposed a test to decide whether a computer program was able to think (Turing, 1950). His proposal was based on a society game in which people try to guess whether they are dealing with a man or a woman, purely based on questions and answers exchanged through pieces of paper. Turing gave this game a little twist: to test whether a particular machine, more precisely a computer program, was intelligent, he proposed a game where you had to guess whether you were interacting with a program or a human person. The program would be deemed intelligent if the judge was unable to differentiate between the two, in other words when the computer program deceived the judge in believing that he or she had been dealing with a human while the interaction had in fact been with a computer program.

Turing unfortunately died too early to make major technical contributions to Artificial Intelligence, which is really unfortunate because he certainly would have been able to do so. So only the Turing test remains as his key legacy with respect to AI. But at the time there was certainly no unanimity on whether this test was a good one. The majority of scientists and developers working on Artificial Intelligence have considered the Turing test to be a sidetrack. For example, Marvin Minsky, one of the founders of AI and Turing award winner, called it "a joke."[5] You

[5] https://www.youtube.com/watch?v=3PdxQbOvAlI from 23:40.

will search in vain for papers on the Turing test in the vast number of papers already published in AI journals or conferences over the past decades, except perhaps to question its utility. Nevertheless, many philosophers and psychologists use the Turing test as their main vehicle to discuss whether or not AI has been making progress. The arrival of ChatGPT only amplified this idea. Many users became convinced that ChatGPT has indeed been able to pass the Turing test because the texts it produces, based on human prompts, often turn out to be coherent and grammatically correct. It has become difficult to distinguish them from human-made texts.

Issues with the Turing Test

On closer examination, we can see two problems with the Turing test:

1. The Turing test is at its core based on deception: it is enough to pretend to exhibit intelligent behavior in order to pass the test. The internals do not matter. In fact, for the kind of Machine Learning that underlies current generative AI, including ChatGPT, the processes by which the trained system reaches conclusions are entirely non-transparent. Machine Learning delivers a black box that cannot be opened, even if we would want to, because system behavior is governed by billions of numerical parameters that have no obvious human-understandable interpretation. So we can only evaluate their intelligence by invoking and externally observing system behavior.

To see that deception is an odd way to track progress in a scientific field, compare this with how other scientific disciplines set challenges for themselves and evaluate whether they are making progress. Let's take biology as an example. Progress in biology is certainly NOT being judged by whether biologists manage to make artificial plants or artificial creatures that deceptively look like real ones. Biology is about understanding the nature and origins of life. True, one branch of biology, called synthetic biology or sometimes "artificial life," uses the same methodology that AI uses to understand mind, namely build artificial systems that exhibit biological functions, such as a growing a cell membrane or kidney dialysis (Langton, 1997). But the objective is not to deceive but to understand how biological functions can be materially realized, why they are important for the survival of organisms, and how these functions could have arisen in evolution.

The goal of AI research, from the very beginning, was to contribute to big scientific questions about mind, such as about the relation between mind and matter, mechanistic explanations of learning, the nature of free will, i.e., the realization of autonomy and agency, among other things (Minsky, 1954). Today this goal is largely overshadowed by the enormous commercial pressures on AI developers to come up with information technologies that can generate the huge profits funders who have invested billions of dollars in AI demand. But it is urgent that the original

scientific goals get back on the agenda, if only to avoid the construction and release of half-baked applications that have a potentially very negative effect on society.

Part of reviving a scientific modus operandi in AI is to find ways to evaluate progress in terms of advances on the fundamental questions AI is trying to help solve. For that the Turing test is not the right way. We must think more deeply about what kind of fundamental questions are at stake and then formulate them in terms of reachable challenges with verifiable outcomes, similar to the way David Hilbert formulated his famous challenges to mathematicians in 1900 (Gray, 2000), van Hemmen and Sejnowski formulated 23 outstanding problems in neuroscience (van Hemmen and Sejnowski, 2006) or Johan Hansson formulated, in 2015, the 10 biggest unsolved problems in physics (Hansson, 2015).

There is a widespread opinion, sadly also among some of those responsible for science policy and its enactment, that AI is not a science and that it is a matter of tinkering until magically something remarkable pops up, like ChatGPT. This is a mistake. ChatGPT rests on a long line of scientific insights and experiments, going back to the 1950s. The limitations of generative AI based on statistical neural learning were already known and discussed in the 1960s, and so were possible remedies developed in the 1970s and 80s. Just like chemical engineering relies on chemistry, or medicine on biology, the engineering of AI requires a solid scientific basis as well.

2. The second problem with the Turing test is that it is easy to deceive us humans because we cannot help but take an anthropomorphic stance when describing and explaining the behavior of complex systems. This stance spontaneously ascribes beliefs, desires and intentions to agents and assumes that they have knowledge and make rational decisions (Dennett, 1987). The anthropomorphic stance serves us well when dealing with other people, but we often also apply it metaphorically to physical entities (particularly in Shamanistic and ritual contexts, Turner, 1974), living entities like pets (McFarland, 2008) or machines, as when saying "my car does not want to start" or "the computer does not understand my request."

The over-adoption of the intentional stance means that it is not a reliable basis for judging whether a particular AI system is intelligent or only appears to be. It is only when we learn more about the internals of a system that we can judge whether a system really deserves to be called intelligent. Often, when we learn how a remarkable result is achieved, we are later disappointed when we find out how it is done, particularly if there is some kind of trick being used that does not resonate with what we believe intelligence to require. For example, many people are surprised that ChatGPT has no access to the meaning of a text although it seems so. Its output is entirely based on predicting the next word in a sentence, taking the context and vast amounts of existing texts, including prompts, into

account. When people become aware of this, they start thinking that ChatGPT is not so intelligent after all.

Disappointment after realizing how an AI system works is also the reason why the definition of what AI is or has accomplished has been shifting throughout the history of the field. Playing chess, solving differential equations, checking the truth of a complex mathematical proof, or scheduling railway traffic for an entire country were all competences considered beyond the reach of machines and requiring human-like intelligence until AI programs turned up that could achieve them – at which point they were no longer considered to require intelligence. This is of course an overreaction in the other direction. It is not because we understand life to be based on biochemistry that living organisms are no longer considered alive.

Measuring Progress in AI

AI scientists are well aware of these two issues with the Turing test. Partly as a response, the machine learning community has molded the Turing test into a more objective methodology, which is more or less standard in engineering. The recent literature is full of papers that first propose a more concrete test (for example, correctly label images or translate a text in another language), then define quantitative measures to objectively establish the level of performance for this test, and finally experimentally and systematically verify how a particular AI technique fares while doing the test. The same test is also given to human subjects so that comparisons can be made to see how far the AI system performs with respect to human subjects.

This quantitative comparative methodology has given an enormous impetus to machine learning research, with tests and datasets being shared and scoreboards announcing on a daily basis which team has the best outcome so far.[6] It has resulted in a race toward "better than human" results for a broad scala of intellectual tasks. Initially dedicated algorithms and training data sets were employed, each tailored to a particular task. But after a sufficient number of successes, the target at the moment is to propose systems that are versatile on many tasks – without tweaking the algorithms each time or introducing new data for training. Even more ambitiously, the ultimate goal is to surpass human intelligence to reach AGI (Artificial General Intelligence), capable of doing any task human intelligence can handle but better (Kurzweil, 2005).

But despite the more objective nature of this methodology, there are still many fundamental flaws, which actually also have been discussed abundantly in the technical literature. The proposed tests turn out to be problematic for a number

[6] https://www.kaggle.com/

of reasons: (i) We are faced with the classical problems that all forms of empirical testing have: representativity of samples, hidden biases, outliers, unforeseen contextual effects. (ii) Intelligence is about dealing with an open world in which there is constant rapid change. Inevitably, test sets quickly get out of date and testing with fixed training and testing sets does not test the extent to which a system can cope with change. (iii) Many aspects, particularly those having to do with meaning, cannot be operationalized for automatic application to large amounts of outcomes and hence approximate measures are used. A good example is in machine translation. Existing measures such as BLEU (Bilingual Evaluation Understudy), METEOR (Metric for Evaluation of Translation with Explicit Ordering), LEPOR (Length-Penalty, Precision, n-gram Position Difference Penalty and Recall), etc. all compare surface features, namely similarity of words and word sequences (n-grams), but not whether the meaning of a source sentence has been captured by the translation. The reason is simple, it is much harder to operationalize meanings and apply an automatic criterion over a large test set.

The consequence of these difficulties is that AI developers cry victory in better-than-human performance for the test sets and performance measures they have but their AI systems nevertheless fail in real world conditions, particularly in open real-world environments where unusual events happen regularly. The mistakes that image recognition systems make for recognizing road traffic signs is a good example (Pavlitska et al., 2023). It is enough that there is some dirt on a traffic sign, a sticker or other light conditions as used in the training set, while a sign may be dramatically mis-categorized with possibly dangerous consequences. This difficulty is one of the reasons self-driving cars are not deemed safe. The failure of IBM's Watson Health system is another illustrative example. Medical diagnosticians are confronted all the time with unusual cases which do not follow a standard pattern and require deeper models and thinking. The unusual cases will hardly show up in the test data. So also in this domain, over-optimism based on high success rates on test data do not pan out in real world conditions (Strickland, 2019).

Turing's Curse

What then is Turing's curse? It is the risk to become obsessed with the published benchmarks and measures and participate in a race to optimize for them, losing sight of the real-world conditions and particularly the changes that are unavoidable in open environments. The obsession has a tendency to lead a community of researchers working on a specific topic in alleys that may turn out to be a cul-de-sac. It precludes spending resources on other approaches that are doing badly on the benchmarks, certainly in the initial stages, but may in fact lead to more profound research results in the long run.

Turing's curse is also the risk to take the comparative methodology too seriously. After an AI system has had a positive and possibly higher outcome for benchmarks,

we are asked to accept that the system is ready for widespread use and worthy of being an adequate replacement for a competent human. We are asked to give that "certified" system responsibility and agency, as we would to a human person, and are asked to trust it. We are told the human expert – the radiologist, architect, teacher, programmer, truck driver, researcher, journalist, artist – is declared to be no longer needed and it is pointless to train or hire new human experts.

If this opinion is obviously outrageous and not shared by everybody working in AI, it is all too common among orthodox members of the machine learning community. A famous illustration is the claim in 2016 by Geoff Hinton (another Turing Award winner) that "People should stop training radiologists now. It's just completely obvious that within five years deep learning is going to do better than radiologists."[7] We are many years beyond the date of Hinton's prediction and – fortunately – there is no sign of radiologists sacked in their jobs, in fact there are too few of them.

Another illustration involves similar claims that programmers will become obsolete in the very near future because large language models appear to be able to do it, as illustrated by ChatGPT (trained on massive datasets of code examples) or GitHub Copilot (Microsoft). At first sight the outcomes are very impressive. These tools provide snippets of code and authoritative sounding explanations, based on a vast training set of human-made code as well as tutorials and textbooks. But unfortunately, there are also blatant mistakes with high security risks, so that programmers are told never to use these tools unless you know yourself how to write the relevant code (Vaidya and Asif, 2023). The mistakes are partly due because of the fact that faulty code was part of the training set and because generative AI will deviate from the most probable pattern to avoid outright copyright theft. The AI-based coding tools can lead to a productivity boost for mundane tasks, but you need to be a highly skilled programmer to use them properly. This is probably a pattern that we will see in most expert domains.

Conclusion

The purpose of this short essay was to argue that the Turing test and the comparative methodology that follows up on it should be taken with a big grain of salt. Consequently, the lack of respect for human expertise that is displayed by AI evangelists is not warranted and neither is the pressure to aggressively spread AI to all corners of society at this point in the development of the technology. AI is certainly an agent of change but there is not enough awareness of its limitations.

[7] https://www.youtube.com/watch?v=2HMPRXstSvQ

References

Gray, Jeremy (2000) *The Hilbert Challenge*. Oxford University Press, Oxford.

Hannson, Johan (2015) The 10 biggest unsolved problems in physics. http://www.diva-portal.org/smash/get/diva2:996740/FULLTEXT01.pdf

Turing, Alan (1950) Computing Machinery and Intelligence. *Mind*, LIX (236): 433-460, doi:10.1093/mind/LIX.236.433

Kurzweil, Ray (2005) *The singularity is near*. Viking, New York.

Langton, Ray (1997) *Artificial Life. An Overview*. The MIT Press, Cambridge Ma.

McFarland, David (2008) *Guilty Robots, Happy Dogs: The Question of Alien Minds*. Oxford University Press, Oxford.

Minsky, Marvin (1954) Matter, mind and models. Updated regularly and reprinted in Minsky, M (ed.) (1968) *Semantic Information Processing*. The MIT Press, Cambridge Ma. https://web.media.mit.edu/~minsky/papers/MatterMindModels.html

Strickland, Eliza (2019) How IBM Watson Overpromised and Underdelivered on AI Health Care. *IEEE Spectrum*. https://spectrum.ieee.org/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care

Pavlitska, Svetlana, Nico Lambing and Marius Zoelner (2023) Adversarial Attacks on Traffic Sign Recognition: A Survey. arXiv:2307.08278 [cs.CV]

Turner, Victor (1974) *Dramas, Fields, and Metaphors: Symbolic Action in Human Society*. Cornell University Press, Cornell, NY.

Vaidya, Jaideep and Hafiz Asif (2023) A Critical Look at AI-Generated Software. *IEEE Spectrum*. https://spectrum.ieee.org/ai-software

van Hemmen, Leo and Terry Sejnowski (2006) *23 Problems in Systems Neuroscience*. Oxford University Press, Oxford.

AI as an Engine of Change of our View on Agency

Johan Wagemans, University of Leuven

In her provocative essay, "AI as an Agent of Change," written for the 2023 KVAB Thinker's Cycle, Helga Nowotny has reflected on several interesting similarities and differences between the printing press and AI as agents of change. With admirable scholarship and eloquence, she analyzes a wide range of historical, social and cultural aspects of these two revolutionary technologies. She presents an impressive synthesis of her findings, draws interesting conclusions from them, and puts them into a breath-takingly broad perspective, with important lessons for governments, policymakers, companies, organizations and individual stakeholders, including scientists and, in a sense, all citizens who are faced with the new developments AI brings them.

In this short commentary, I will focus on the notion of agency, which is at the core of the analysis. To what extent can we really attribute agency to AI? Helga Nowotny connects the agency of machines to its functions and intentions: "machines are built to fulfill certain functions. They have human intentions inscribed into them." She then points out that humans have a tendency to attribute agency to machines, based on "a deeply rooted anthropomorphic tendency to view the behavior of another entity or object in terms of mental properties." This remains "relatively harmless if it refers to familiar technologies. … However, when it comes to AI it can transform … into a dangerously compelling illusion of being in the presence of a thinking creature like ourselves." In her book *In AI We Trust*, Helga has identified an interesting paradox in this regard: "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."

In the remainder of this commentary, I will try to clarify that knowledge of the factors that drive human behavior or underly our preferences has always been exploited, and that there is, in fact, a continuum of losing control over our own behavior. AI is not so new in this regard, it just does it better, and it does it in less transparent ways. When we walk along the shelves of a supermarket, looking for an item we need, we might not be aware that the most expensive products are stalled at eye-height, while the cheaper brands are either placed higher or lower, and thus more difficult to reach. This economical behavior of the supermarket staff, who want to make more profit, makes perfect use of scientific insights into aspects driving the behavior of the average consumer (e.g., limited attention span, normal perceptual processes, automatic decision-making with little cognitive control). Search engines make use of our past search behavior to

sort hits for us, and additional economical forces (mainly companies paying to get their website higher) will also play a significant role, perhaps beyond the user's own conscious awareness and lowering their control. Recommender systems on Netflix, Spotify, Instagram and the like go one step further still, although many consumers might actually like this feat.

With regard to the specific case of aesthetics of images, all cameras in our smartphones now use built-in software to make better pictures, and Generative Adversarial Networks can be used to enhance the aesthetic qualities of pictures in some kind of black-box post-photography editing stage, based on machine learning networks trained on tens of thousands of images. Vision scientists can examine the parameters that are modified and try to understand factors driving aesthetic appreciation, such as enhanced contrast, color, sharpness, depth, etc., but understanding these factors does not mean that we are given control over what we do with the images.

In a more recent development, called computational aesthetics, machine learning models are trained to predict human aesthetic preferences for images. So far, this has not been very successful yet, although the big tech companies invest a lot of money and research time into this, partly because the quality of the training data is limited, and aesthetic preference is far from universal. Instead, culturally, or socially co-determined influences, as well as very personal factors, play major roles too. Should we be afraid for such developments? As long as we are aware of the goals of the systems we are using, and as long as we understand at least a little bit about the essence of machine learning based on previously acquired data of other human users, we are still able to keep control of our own behavior. We can use these systems without "handing over control" to them. I am personally engaged in a large project in which we aim to predict, explain and understand human aesthetic preference for images. We start from existing machine learning models, but we will enrich them with our knowledge about human perception, including the central role played by perceptual organization in interaction with memory, emotion, expertise, personality and so forth. Our goal is not to compete with the big tech companies, but to make significant scientific advancements. With this knowledge we can then inform the public about all the factors that get into the mix, and how they can enrich their aesthetic experiences by exploiting such knowledge.

I believe this resonates with Helga Nowotny's plea "to move away from the simplistic binary utopian-dystopian scheme of thought" and her emphasis on "the important role which science has to play in conveying to the public how AI works." For me, AI is still an engine, not a real agent, because its apparent intentionality and agency is always initiated by humans: those that intentionally develop and train the neural networks to perform a specific function, usually narrowly defined and almost always based on training data that are unintentionally delivered by thousands if not millions of humans.

4. Reactions from Policymakers

How Is Flanders Doing in AI?

Bart De Moor, KU Leuven

Artificial Intelligence (AI) – or, to use a more appropriate term, *assisted intelligence* – was recently described, in an impressive and exhaustive report[8] by the Dutch WRR (*Wetenschappelijke Raad voor het Regeringsbeleid)*, as the "New System Technology."

Indeed, the subsequent industrial revolutions of the last three hundred years were all characterized by the introduction of new "system technologies," the impact of which would drastically change global society. Think of energy production and consumption (coal and steam, fossil fuels like petroleum, electricity, nuclear power plants, etc.), the mechanization, roboticization and automation of manufacturing industries and industrial processes, the revolutions in mobility (trains, cars, airplanes, etc.), global informatization (computers) and the revolutions in communication technologies (audio-visual media, world wide web, internet).

All of these are examples of *system technologies* that, once they conquered the world, were there to stay. All these system technologies build on previous generations: the world wide web and the internet build on our world-wide communication systems, on computers (Moore's law), on (software) automation, and they would obviously be impossible without energy provision.

AI, as a new system technology, adds a new layer to this evolution, building on information sciences (including mathematics, software engineering, etc.) and technologies (computers and servers), communication (worldwide connectivity) and data sharing (e.g., wireless interactions, websites, databases, etc.), requiring vast and massive demands of power and energy (which all too often we take for granted). The tsunami of sensors and data, sometimes called *the new gold,* drives an increasing number of data-driven opportunities, services and applications in scientific research, health and medicine, industry, mobility, government administration and our daily lives.

The advent of new system technologies typically comes with (or is based on) new scientific insights and technological breakthroughs. They result in novel applications, new business models, new benefits and, unavoidably, nuisances and risks (sometimes unforeseen) for citizens and society. Because of the ubiquitous

[8] https://english.wrr.nl/publications/reports/2023/01/31/mission-ai.-the-new-system-technology

impact on all dimensions of our daily lives, governments on a global, national and regional level have no choice but to get involved in the development of new regulatory legislation, in the approval of worldwide protocols, in the deployment of infrastructure, in the mitigation of risks, in the protection of institutions, companies, citizens and all other stakeholders, so that they can all optimally benefit from the new technologies. And of course, often these new technologies will provide opportunities to raise new taxes. For sure, the new system technology of AI is no exception to all of these features.

All these system technologies typically cause non-technical deficiencies in the way our societies operate. There are *democratic* deficiencies, because decision-makers (e.g., in parliaments) do not always understand the ins and outs of these new technologies, as well as how their benefits and threats could impact society. Often this leads to overregulation, which can slow down innovation. There are *legal* deficits, because parliaments and governments are too slow in formulating and voting adequate laws and sometimes they will only do so after unfortunate incidents or when they are surprised if not overwhelmed by global evolutions. Finally, there are *ethical deficits*. They originate not in the "*how* of new science and technology," but in the "*what*." Not in *how* new AI applications are too be implemented, but rather *what* the impact and consequences, both foreseen and unforeseen, could possibly be. Ethics deals with necessary choices that must be made within a whole spectrum of available possibilities facilitated by progress in science and technology.

Flanders, as one of the three regions in Belgium, is no exception when dealing with this global impact of the new AI system technology. Five years ago, the Flemish Government decided to launch an ambitious Flanders AI Program[9] of 32 Mio €/year, which comprises three pillars: a *research program*[10] of 12 Mio €/year, *an additional budget* (15 Mio €/year) *for R&D grants*[11] *to companies* that develop innovative AI applications, and 5 Mio €/year for so-called "supporting measures": a *Knowledge Centre Data and Society*,[12] the *Flanders AI Academy*[13] and several communication initiatives, including the *AI citizen science initiative AMAI.*[14]

FAIR (Flanders AI Research Program) includes the five universities in Flanders (Leuven, Gent, Antwerp, Brussels and Hasselt), and the four strategic research centers (imec (nanotechnology), VIB (biotechnology), VITO (environment, energy

[9] https://www.ewi-vlaanderen.be/en/flemish-ai-plan/broad-context

[10] https://www.flandersairesearch.be/en

[11] https://www.vlaio.be/nl/begeleiding-advies/digitalisering/artificiele-intelligentie

[12] https://data-en-maatschappij.ai/en/

[13] https://www.vaia.be/en/?lang=en

[14] https://amai.vlaanderen

and sustainability) and Flanders Make (manufacturing)). It involves several dozens of professors and principal investigators, and several hundreds of PhD students and post-docs, in over more than forty research groups.

Just like in its first edition (FAIR 1.0, 2019-2023), the research activities are organized in so-called *Grand Challenges* on the one hand and *Use Cases* on the other hand. For the second 5-year period (FAIR 2.0, 2024-2028), there are two Grand Challenges. The first Grand Challenge (the larger of the two) is about *AI Driven Data Science* and endeavors *to support complex decision-making and the creation of actionable insights from the exploitation of ubiquitous data.* The objectives of the second Grand Challenge, on *Situated AI,* are in *supporting complex task execution in a dynamic environment with (semi-)autonomous AI Systems, collaborating in real time with each other and with people.* The two Grand Challenges share common translational values and objectives with well-defined criteria for AI that is human-centered responsible, resilient, performant, data-efficient and sustainable from the energy point of view.

While designing FAIR 1.0, we discovered an apparent paradox when interacting with potential users of AI technology. All of them, without exception, acknowledge a pressing demand for more AI in their biotope. But AI is a "container term" that covers a wide diversity of problems and solutions. It is very important to manage the "AI expectation," because often potential users expect miracle solutions that current-day AI simply cannot offer (and in many cases will never be able to do so even in the future). When asked to describe their needs in more specific terms, potential users often do not know, let alone understand, the typical AI jargon to really pinpoint the exact problem for which they are looking at an AI solution. In other words, the paradox we discovered implies that there is *a broad demand for more AI* on the one hand, while, on the other hand, there is an absolute semantic threshold when trying to formulate or "articulate" in detail the questions at stake. This is the paradox of *demand articulation.*

Several measures were taken to fill this gap.

VAIA, the Flanders AI Academy, organizes several hundreds of lifelong learning activities on AI per year, in which the resolution of the demand articulation paradox is paramount.

VLAIO, the agency that subsidizes AI R&D in companies, also started organizing a lot of customized information trajectories, providing a lot of best practices.

In the research program, we have launched more than thirty use cases, that can serve as a role model and inspire future potential users. These use cases are grouped into four clusters: Health, Planet and Energy, Industry and Society. In the Health cluster, the use cases are *monitoring at home, real-time real-world*

biomedical monitoring, medical imaging, single cell molecular biology, digital twin cardiology, AI in intensive care, personalized dermatology, sports monitoring. In the Planet and Energy cluster, we have *natural environment monitoring, geo-urban platforms, and AI for the smart grid in Flanders.* In the Industry cluster: *smart machines, monitoring and control of production machines, product optimization, straight-through-digitization manufacturing, prognostics and health management for assets and refurbishment.* Finally, in the Society Cluster, we have *AI to optimize public employment initiatives, digital humanities and AI for education and training.*

All these initiatives comprise the ambitious plans for FAIR 2.0, the second phase of the Flanders research program for 2024-2028. A minor point is that, against all expectations, the current Flanders Government decided not to increase the budgets for the overall AI program. At least for 2024, the budget remains *status quo,* despite an extremely positive assessment of FAIR 1.0, the research program 2019-2024, by our international Scientific Advisory Board and, in addition, by an independent committee of international experts. This is of course deplorable, as in all regions and countries that neighbor Flanders, R&D budgets for AI have considerably grown and will continue to do so in the short term, picking up on international trends as sustained in the US and China. Therefore, all stakeholders in Flanders sincerely hope that as of 2024 the prospective Flanders Government will catch up and synchronize with these international trends.

Summary Speech for the "AI as an Agent of Change" Event

Lucilla Sioli, AI and Digital Industry, European Commission

As the Director of the "Artificial Intelligence and Digital Industry" directorate of DG CONNECT, in the European Commission, my portfolio encompasses a broad spectrum of responsibilities. These include formulating AI policies in support of the development of AI in the EU. This pertains to promoting research, innovation and deployment of AI, making sure that we have sufficient talent in the EU, but also to the regulatory framework on the use of AI, the AI Act, stimulating the development and the uptake of trustworthy AI. Additionally, I oversee the governance of AI, which includes the promotion of trustworthy AI internationally.

In discussing the crucial topic of AI's role as an agent of societal change, I wish to draw your attention to the insightful recommendations provided by Helga Nowotny in her report, designed to foster a more informed, humanistic and balanced approach to AI development and its integration into society. In the following, I would like to emphasize how these recommendations resonate with the key policy and regulatory frameworks of the European Commission (EC), namely the Coordinated plan on AI and, respectively, the AI Act.

In her first recommendation, Nowotny highlights the need for a large public campaign to educate citizens about AI, its diverse applications and its inherent limitations on the basis of a digital humanism perspective. This aligns fairly well with the EC's coordinated plan, which emphasizes a human-centric approach and the need to boost digital skills and AI literacy among citizens. It also harmonizes with the AI Act's focus on transparency and accountability, ensuring citizens understand and trust AI systems. As part of the actions planned, the Commission, via the Digital Education Action Plan 2021-2027, supports traineeships in digital areas, with a focus on AI skills, and the development of ethical guidelines for AI and data usage in teaching and learning for educators. Additionally, under the plan, the Member States are encouraged to refine and implement the skills dimension in their national AI strategies, in collaboration with social partners. This involves promoting computational thinking, creating AI educational programs, increasing AI training availability and supporting effective AI educational technologies.

Nowotny's second recommendation advocates prioritizing basic, ERC-style research in AI, suggesting a focus on humanistic aspects, as well as addressing the imbalance between public and private AI research funding. Concerning the need for research and innovation in AI, the EC's coordinated plan outlines several key initiatives. These include the establishment of European Partnerships in Horizon Europe for driving innovation in AI, data and robotics, focusing on a human-centric and trustworthy European vision of AI. Further, in a top-down approach, the plan seeks to strengthen the EU's excellence in AI research and innovation by providing funding for the development of the next generation of AI systems, which will be greener, more autonomous, transparent and non-biased. The plan also introduces the AI Networks of Excellence Initiative to foster a strong European research alliance. Additionally, the policy framework promotes trust in AI systems, ethical AI development, and multidisciplinary research. Finally, it encourages the acceleration of private and public investments by leveraging EU funding available through programs like Digital Europe (DEP), Horizon Europe (HE), and the Recovery and Resilience Facility (RRF), for example setting up facilities that facilitate testing and experimentation.

Nowotny's final recommendation calls for robust support for research into AI's impact on society, particularly in areas not likely to be pursued by large corporations, to understand AI's beneficial applications and avoiding social harm. It also stresses the need for interdisciplinary approaches in understanding and applying AI. This echoes the AI Act's commitment to ensuring AI's adherence to fundamental rights and values. Indeed, the AI Act follows a risk-based approach whereby rules target AI systems used in contexts where people's fundamental rights and safety are at risk. Moreover, as already discussed, the Coordinated Plan promotes the development and uptake of human-centric, trustworthy, secure and sustainable AI technologies.

In summary, Nowotny's recommendations align well with the specifications of the Coordinated Plan on AI of the EC and the AI Act, with some nuanced differences. I therefore welcome Nowotny's recommendations, as they offer a valuable perspective that complements and enriches the existing plans and acts.

Finally, regarding the future perspectives, we are preparing several measures, from basic research to deployment and financing instruments, in order to put Europe on the map of Generative AI.

5. Reports of Stakeholder Workshops

Stakeholder Workshop I – 12 September 2023[15]

Theme: AI as an agent of change: How are AI/ChatGPT used, experienced and supported in the formal education as well as in the information and training of the broader public in Flanders?

The thinker and the steering group have distributed in advance several questions to the participants.

- What are the current uses of AI/ChatGPT in and outside the classroom and for what purpose?
- What has been your and your students experience so far?
- Is there curiosity and enthusiastic adoption?
- Which limitations and problems did you encounter?
- What is the role of teachers? How can they help?
- Do teachers feel left alone? Do they need support from other disciplines?
- Are institutional guidelines sufficient?
- What else is needed to enable productive use for all?
- Where would you like to be in three years?
- Explore novel uses, questions, outreach: do you have an example?
- How do you foresee to navigate between positive and negative future visions?

One important observation shared among the participants is that there is already a widespread and diverse set of actions and initiatives dealing with the topic of generative AI for teaching and research and communication toward society and the general public (see appendix[16]).

Generative AI is already widely used in Flanders in education at the primary, secondary and university level, as well as in leisure activities and in professional environments. We can list here examples of use like language correction, generation of reports, review and summarizing of documents and research papers, as a search engine etc. In general, the public is aware that these services are fairly

[15] Participants: Helga Nowotny (thinker), Charlotte Vandooren (imec/RVO society), Johan Suykens (Engineering, Master of AI program chair, KU Leuven), Cynthia Van Hee (linguistics, UGent), Els Lefever (linguistics, UGent), Hugo De Man (Engineering, KU Leuven, imec, KVAB), Joos Vandewalle (Engineering, KULeuven, coordinator, KVAB).

[16] Appendix Initiatives and actions on the use of AI/ChatGPT in Flanders for education, training and information outside the research community: Aanbod voor onderwijs (amai.vlaanderen), AI voor Vlaanderen, AI4Belgium, Scivil, Kennis centrum data en maatschappij, Digisoc, amai! Imec-vub smit, imec-rug IDlab, imec-ua IDlab, citip-kul, rvo-society brightlab, nerdland, citizen projects, ChatGPT for schools pptx, FARI (AI for common good) vub -ulb.11-12sept.

new and to be developed further, and that they are likely to stay with us in the future.

The stakeholders agree that generative AI and general AI services are not magic, but that they mainly involve mathematical optimization of predictive models that are trained with massive compute power and rely on massive sets of data. Moreover, AI consumes massive amounts of energy, the data sets employed may be unreliable and biased, while there tends to be little respect for the privacy and the authorship of texts. Stakeholders also feel, however, that the public and the younger generations are not sufficiently aware of these various aspects.

The big companies are in control of the data and the media, and the individual user has no insight and understanding of the process, value, limitations and validity, and hence there is often a false belief of trustworthiness. The big companies have launched this technology on the public prematurely, with the goals of making profit. Universities and public research centers do not have the same number of resources, and therefore they must limit their effort to smaller and more focused themes that are handled with better objectivity and honesty. Yet numerous research questions are still left open such as the explainability of generative AI, the correctness proofs, hallucinations etc. If we let them go, AI will outsource produced knowledge in uncontrollable and unexplainable ways. The public may be insufficiently aware that it does not communicate with a human being but with a machine that uses uncontrollable and often untrustworthy massive data sets.

This is why a new form of science communication is needed. Scientists must create a dialogue with citizens to explain that science is a practice of organized skepticism. Rather than telling the public that refusal of scientific evidence is the same as a lack of understanding, scientists should explain how science works by interaction and experiment and in this way that it can be of help in their life, job, and education. Public engagement of science is gaining importance. Science has a moral duty to avoid that people are disoriented.

The government, industry, education, media and arts are aware and interested in these new services and are willing to act and engage. So, the topic is open for public discourse in which ordinary citizens of all ages are eager to participate.

Master AI students are enthusiastic about using AI, e.g., to get an initial idea about a subject, to let it correct a text, or to let it generate programming code. They are also aware of the limitations. Since ChatGPT is often not reliable on factual information, one should further debug the programs, one should further check references on correctness and existence, etc. ChatGPT is not yet a reliable tool, though it is having already a number of impressive features. Moreover, it cannot be used with sensitive or personal data due to privacy issues and data or subjects under non-disclosure agreement. With the Master AI program at KU

Leuven a template file has been created that has to be completed by the students for all reporting in course assignments and for the master thesis. It also contains a code of conduct related to all possible forms of uses of ChatGPT and other AI writing assistance tools. It also explains how to comply with exam regulations. At KU Leuven guidelines for responsible use of generative AI tools in research and education have been developed, where the code of conduct proposed by the Master AI program has also been taken into account.[17]

The quality of tools like ChatGPT should improve. Especially factual information and references should be correct and reliable. If factual information is correctly implemented, it will become a very important and disruptive tool, making possible many exciting new things in the future. At the same time, however, new AI tools are also likely to introduce new and unforeseen problems, and this makes it hard to predict at this point how this technology will develop.

The education and training of language teachers and translators is rather deeply affected already. Of course, it should include the responsible and inclusive use of these digital services and prepare for wider use. Also, the limitations of the services and the essential added value of human knowledge, sentiments, emotions and wisdom should be explained. In other words, the prospects for professionals in these fields are changing, but not necessarily in a negative direction.

The role of engineers, data scientists and designers of these services is changing as well, and this should be reflected in their education. They can no longer do their work in isolation from the public or the experts in the humanities and social sciences. There is currently a growing awareness and willingness to act in this respect, both internationally and in Flanders.[18]

Although the participants have a diverse background, activity scene and career path, there is a common understanding and conviction that the themes discussed include several important insights, issues and attention points that need a comprehensive approach in Flanders both for the formal education from primary and secondary schools to universities and for the general public in their leisure and day-to-day activities. Such a plan should be institutionalized and work with experts on the ground targeting different audiences in order to bring about a deeper understanding, a demystification of the issues involved, a balanced view

---

[17] https://nieuws.kuleuven.be/en/content/2023/ku-leuven-opts-for-responsible-use-of-generative-artificial-intelligence-in-researchand-education

https://www.kuleuven.be/english/education/student/educational-tools/generative-artificial-intelligence

https://research.kuleuven.be/en/integrity-ethics/integrity/practices/genAI

[18] Putting the common good at the heart of AI, data and robotics research, https://www.fari.brussels/nl

on benefits and a critical attitude. In this way the people will serve as the agents of change.

Stakeholder Workshop II – 15 September 2023[19]

Theme: AI as an agent of change: What is the multidisciplinary experience of the Flemish researchers, that are active on AI basic research and applications, with respect to the societal aspects of AI, and more specifically ChatGPT?

The thinker and the steering group have distributed in advance several questions to the participants. The participants first gave some remarks on the general theme.

"Responsible AI" is already an important part of the Flanders AI[20] initiative, where several participants are involved. Flanders AI has an extensive program covering strategic basic research into artificial intelligence, based on the needs and demands of companies, organizations, the government and citizens. Clearly most participants are aware of the main societal issues, but the public is not so much aware. Some issues can be taken into account in the fundamental research and design phase of the machine learning methods, while most should be addressed within the specific application context when applying the machine learning methods into services and products.

1 - How are you dealing with the problematic issues of AI (bias, explainability, accountability, transparency, fairness)?

Overall, participants expressed AI as beneficial for science, with many possibilities for its use in their research practices. Especially the use of generative AI was considered a game-changer for science. However, all acknowledged that important issues and risks of AI should be taken into account. Problematic issues of AI are bias, explainability, accountability, transparency and fairness.

Bias and fairness are strongly related, while, similarly, explainability is linked to transparency in AI. Explainability/transparency is a precondition for accountability and fairness, but also for confidence/trust and empowerment of end-users

[19] Participants: Helga Nowotny (thinker), Sabine Demey (imec/FlandersAI), Tony Belpaeme (Engineering, UGent), Tijl DeBie (Engineering, UGent), Stein Aerts (Medicine, VIB, KU Leuven), Sien Moens (Computer Science, KU Leuven), Spyns Peter (Dept EWI, Vlaanderen), Tias Guns (KU Leuven), Thomas Demeester (UGent), Pedro Goncalves (nerf), Guillermo Perez (UAntwerp), Hugo DeMan (Engineering, KU Leuven, imec, KVAB), Joos Vandewalle (Engineering, KULeuven, coordinator, KVAB), Ine Van Hoyweghen (KU Leuven, coordinator, KVAB), written answers by Johan Suykens (Engineering, KU Leuven).

[20] https://www.flandersairesearch.be/en

and consumers. In order to rank them according to importance, explainability/ transparency is considered the highest, as it is in a sense more fundamental.

Aspects of bias in decision-making can be related to the bias term in a classifier, diversity sampling, and others. Much research has been done on explainable AI. One also aims at reproducible research and open-source software, resulting in better transparency. Concerning the accountability of AI, a distinction is needed between general methodological machine learning research versus AI in a specific application context. Although one may have an intuitive opinion on fairness about what is right or wrong, this will depend on the application context and could prove hard to quantify in an objective manner. Moreover, because definitions of fairness vary between different disciplines (e.g. computer science, law, economics, philosophy, sociology), there is a need for transdisciplinary research on fairness in AI.

There is active research in Flanders on several of these issues. One researcher explicitly deals with explainability and transparency. Within the field of explainable AI, there is a group of experts in formal methods who approach this algorithmically from two fronts. First, they develop alternative/modified AI techniques that provide explanations or enable efficient checking of possible explanations. The second approach proposed is that of taking a vanilla AI model and checking that it conforms to some specification in a language that is considered (by experts) as explainable, which is related to verification of computer systems. Another member works mostly on transparency of generative models: building additional control in representations and neural architectures (e.g., in text to image synthesis/ diffusion). Other members in Flanders conduct research on causal models, causal relations and logic reasoning. Overall, participants expressed the need for basic research into generative AI.

- Which issues are most problematic for you and why?

In particular explainability and fairness are difficult issues to tackle: often one will see an accuracy versus explainability trade-off. Fairness is usually a subjective notion. Another observation is that the data and their distributions on which our current large foundation models are trained are not known. Moreover, there is ignorance of all details of the models used for training the foundation models. Regulators and scientists should have access to the inner workings of these models – how they were trained and on which datasets. If these models are transformed into closed products, they will be unavailable for thorough inspection, replication and testing. Quantifying accountability may be the hardest issue since it is closely related to legal matters. More generally, is the programmer, the company, the software, or the data responsible if something goes wrong due to a bug or attack? Adding AI to the possible culprits only renders this concern more complex as there's the training of the data, the designer of the AI algorithm, the person who chose the AI framework, and so on.

Participants also mention the broader societal impact of AI. For example, discussion was raised on the safety of My AI on Snapchat. Chatbots with ChatGPT were considered as worrisome in particular, notably for children, because it is prone to propaganda and cyberattacks.

While some members are optimistic about technology in general and AI in particular, and thereby refer to overregulation as it has happened with the genetically modified food, most members consider regulation as necessary in order to avoid societal harm. The European Commission made a proposal for an EU regulatory framework on artificial intelligence (AI) in April 2021.[21] The proposed legal framework focuses on the specific application of AI systems and associated risks. The Commission proposes to establish a technology-neutral definition of AI systems in EU law and to lay down a classification for AI systems with different requirements and obligations tailored on a "risk-based approach." Some AI systems presenting "unacceptable" risks will possibly be prohibited.

- If we agree on the necessity of regulation, which technical hurdles need to be overcome?

It is relatively straightforward to define a "wish-list" of desirable properties of AI systems (such as on bias, explainability, accountability, transparency, fairness). However, there may exist fundamental theoretical limitations of what is achievable (e.g., trade-offs on robustness versus accuracy, performance versus explainability, transparency versus company's competitiveness). Talking in very general terms about AI systems is often difficult, as one finally needs to take into account the particular application context. Given that one needs to consider the AI system, the data, the application context, the designer, the user and the goal of the AI system, it will often be an interdisciplinary and transdisciplinary task to realize and implement the AI system. In most cases, expertise will be needed from several different fields.

Over the course of the past years, various notions of explainability/transparency and bias/fairness have been formalized mathematically, which is a requirement to be able to impose them on AI algorithms and systems. There is a large gap, however, between these mathematical notions and legal perspectives on the same concepts. Even in the simplest AI-setup of binary classification, fairness can be formalized in multiple incompatible ways, and the differences between them, even if they truly matter to individuals and society, are hard to convey to non-technical people such as business decision makers and legal or ethical advisors. The models change very rapidly; we cannot afford just to say they are black boxes, because

[21] Artificial Intelligence Act https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

they are not. We need experts that understand the computations, optimization processes and keep up with the latest evaluations in technology, and we need access by these scientists to the training of the models. Several participants feel that regulation is unlikely to work and compare it with the regulation on cookies, which actually failed to limit the widespread use of cookies.

One participant argued that governments should refrain from regulating computer systems or AI in particular because national /and international legislation moves too slowly to keep up with AI developments. Instead, encouraging companies to establish standard processes to validate and verify AI-enabled systems should be prioritized. See, for instance, the verification and validation plan of NASA technology (from their systems engineering handbook available online). This should make the discussion more concrete.

Reflecting on this problem, the draft EU AI Act has multiple provisions around explainability/ transparency as well as bias/fairness (and more), but they remain vague from the perspective of computer scientists and AI system developers. The question of how these vague requirements can be translated into concrete implementations remains unanswered.

- What should be done to address these hurdles? What is the role of companies and the role of individual researchers?

The requirements linked to transparency are currently different for universities and companies, which is causing a mismatch, in particular because fundamental AI research is increasingly done in (big) companies as well.

AI developments are in an acceleration phase. Moreover, AI needs a stable environment in order to be able to flourish. However, the geopolitical status of the world is unstable, which is causing fundamental problems. Therefore, achieving a worldwide international agreement on common principles will become essential.

Training and education of engineers is needed. Without educating AI experts with the abovementioned skills, we will lack company employees and researchers with these skills. More generally we need to train students in ICT, engineering, mathematics and statistics to understand the inner workings of the neural architectures used (used in natural language processing, computer vision, speech, etc.), to understand the underlying mathematics and optimization processes. This is considered a moral duty. Moreover, these experts need training on expertise from other disciplines (law, sociology, ethics). Apart from this, these critical skills are a basis for innovation and are absolutely needed in the fast-changing AI landscape. Entirely novel methods and tools are needed for bridging this gap between technical approaches and social/legal/ethical perspectives, aimed at involving all stakeholders with widely varying backgrounds and interests. Such

tools should help in the design of an approach to those issues that is broadly supported, guaranteed legally compliant and technically achievable.

The participants believe that there is an important role for research to contribute to this challenge. Collaborative research across disciplines will be crucial for this, pertaining to, for instance, applications, AI foundations, formalization of fairness/bias, user interfaces, law and ethics, sociology.

2 - Many partly overlapping ethical guidelines exist, but often are under-specified.

- What is most lacking to monitor and implement them? Are sanctions needed?

What is typical for guidelines and regulations is that often they are vague or underspecified to cover many situations, in particular future ones. This is why they should be evaluated and interpreted based on the latest technological advances. If sanctions imply that certain technologies are forbidden, this might hamper technological progress. At this point one should mainly focus on the high risk AI applications as defined, e.g., in EU AI act proposal https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

It is also best to monitor it for each application sector, taking into account the specific application context.

Despite the many remaining challenges, the research community is starting to get a handle on issues related to transparency/explainability, bias/fairness, privacy, etc. In contrast, insufficient attention is dedicated to the broader societal, pedagogical and psychological effects of AI. What is most lacking, according to another member, is guidelines and regulation around AI social companion chatbots, and the use of conversational AI in the form of social media bots. In terms of regulating such applications of AI, one participant believes that the risk-based approach to non-generative AI in the draft EU AI Act is very sensible. Unfortunately, there is also fear that the amendments by the Council of the EU and particularly by the European Parliament, which are aimed at generative AI techniques and foundation models that underlie these applications and which thus deviate from the risk-based approach, come with the risk of creating unnecessary obstacles to innovation in generative AI and foundation models, while at the same time failing adequately to recognize and regulate the high-risk applications thereof. In particular, the risks in terms of large-scale disinformation and manipulation are likely to be underestimated, particularly in applications where people are likely to form emotional connections with AI systems, such as AI social companion chatbots.

- What prevents monitoring and implementation?

As AI accelerates, new developments and future AI systems and services are hard to predict. There is a lack of public and political awareness of potential risks, as they appear to be more subtle than, e.g., direct discrimination or privacy invasion in areas that are recognized as high-risk in the draft AI Act.

- Beyond ethics: If we want to take the likely societal impact into account, how can this best be done?

Participants stressed the relevance of research into the societal impact of AI, with respect to the current research ongoing in this field in Flanders. At minimum, a public and political debate on the role we want to give human-like AI systems in our society is required. Caution is especially warranted toward minors, particularly with AI systems that are likely to result in emotional attachment, let alone systems marketed for such purposes. These should be regulated at the level of the application (as was the case in the original draft), rather than at the level of the technology (as is done in the latest amendments, at least for generative AI and foundation models).

Serious impact on the job market can be expected. Several kinds of jobs might decrease in volume or disappear, while others may need to be adjusted. At the same time, AI will continue to offer new possibilities for jobs in many disciplines.

3 - As researchers we have the responsibility as well to instill critical thinking, digital literacy, trust and confidence in the younger generation and in the public.

- How can we have a democratic debate on AI governance?

For the general public it does not make sense to ask these generic questions, as this invites meaningless discussions which only serve to polarize the population. Instead, debates should be centered around concrete questions and practices of AI. It is best, then, to have a democratic debate on specific high impact AI topics, e.g., ChatGPT, self-driving cars, AI health applications etc., discussing the pros and cons of new developments. Ensure that a cross-section of society and of the scientific community is involved in this debate, meaning that the debate should be informed by AI experts, but not dominated by them. This will ensure the broader societal context, and that the needs and concerns of all members of society are taken into account. We should avoid the mistakes we have seen during the pandemic in this regard.

Also, it is important to educate technological experts and experts in other disciplines about this fast-changing field.

We should make the public aware of the value of the data that they create. The data that we as a public give to the big tech companies might be worth more than the services that the big tech companies give in return to the public. We might think of other business models.

Recently the PhD students in AI in Flanders showed much interest in the societal aspects of AI during a study day. In the engineering faculties of the Flemish universities, students would like to have more training in societal aspects of AI. This type of training should have a trickle-down effect on society. Citizens must learn how chatbots are powered by neural networks and what constitutes their training – mechanisms that demand critical thinking beyond mere digital literacy.

- Can AI become a public good? Under which circumstances?

Most participants hope that this will happen. Academia together with the Open-Source community should play a crucial role in this. When foundation models could become a public good, this is likely to benefit the climate since the training of these models needs a lot of energy. European academic institutions could give incentives to build such public models.

AI is already a public good as most developments are described in paper preprints posted online to be accessed free of charge, as a form of open science. What is not a public good is the detailed knowledge on how to train them and use them. More importantly, the data used to train them is not a public good either. Besides bringing programming as a compulsory element of high school education, the very basics of using AI techniques should probably also become compulsory as they will become a standard tool in programming and computer engineering in general.

- What is the role of interdisciplinarity in all of this?

All are convinced that deep interdisciplinarity is not only important but also vital during the different stages of the AI research and development. Different levels of understanding the technologies can be considered, and not all researchers need the same depth. On the one hand, there are the computer and information and data sciences and mathematical foundations, and, on the other hand, there are several disciplines in the humanities and social sciences – like sociology, law, psychology, education, linguistics, anthropology – as well as the interaction with society and citizens.

6. Conclusions and Recommendations of the Thinker's Cycle

Over a period of nearly one year, this Thinker's Cycle has managed to generate interesting reflections, new discussions and more sustained awareness – in Flanders as well as within the KVAB – on the important role of AI in our society and in numerous scientific activities. The well-known interdisciplinary and visionary thinker Prof. Helga Nowotny greatly stimulated these various activities through her contributions. In her study "In AI We Trust" from 2021, she covers revealing historical parallelisms and presents important interdisciplinary insights. Together with the recent progress and widespread use of generative AI, this has produced a deeper understanding of the ongoing transformation in science and society today. There was a clear convergence among the participants on the urgent need for joint actions and intense cooperation between experts in AI science and technology and scholars in the social sciences and humanities. It is clear that general AI and generative AI offer many new opportunities to users and scientists alike, but at the same time they come with a range of potential risks and social harms, such as bias, unwanted profiling/ discrimination, data misuse and rising social inequalities.

Moreover, both in science and among the users there is a need for better information, more insight and more reflection on ethical conduct. It is important, for instance, to foster the awareness that AI tools are not magic, but that they are produced by mathematical optimization using massive computer power and built on huge data sets. On all these fronts there are major concerns. The mathematics of optimization and the algorithms are solid, but they do not provide a justification of or explanation for the AI products as such. The computer power and big data sets are in the hands of near monopolies worldwide. Moreover, the power that computers need today to update these models is already at the level of a good-sized country's energy consumption. In other words, the various data sets come with a price tag, and, worse, they may contain misinformation that can lead to fake outcomes.

In this context, experts, stakeholders, AI practitioners and the steering group discussed and commented on an inspiring text by Helga Nowotny distributed in advance. It is fair to say that there was good resonance between all participants as described in the reports on the stakeholder meetings, while the various experts contributed inspiring reflections (see Chapter 5). Based on the various discussions, a set of three specific recommendations was formulated.

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 human resources, initially for a period of three years, and potentially 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 interdisciplinarity, and even trans-disciplinarity, that is urgently needed.

Appendix 1 – CV of the Thinker

Helga Nowotny is Professor Emerita of Science and Technology Studies, ETH Zurich. She is also a founding member and former President of the European Research Council.

She has held teaching and research positions at universities and research institutions in several countries in Europe, and she continues to be actively engaged in research and innovation policies at the European and international level. Currently, she is a member of the Board of Trustees of the Falling Walls Foundation, Berlin; Vice-President of the Lindau Nobel Laureate Meetings; Senior Fellow at the School of Transnational Governance, EUI, Florence; member of the Council IeA de Paris; member of the Austrian Council for Research and Technology Development; and Chair of the Scientific Advisory Board of the Complexity Science Hub, Vienna). She was Visiting Professor at Nanyang Technological University, Singapore. She received multiple honorary doctorates, including from the University of Oxford and the Weizmann Institute of Science in Israel.

She has published widely on science and technology studies (STS) and on social time. Her latest publication, *In AI We Trust. Power, Illusion and Control of Predictive Algorithms*, was published by Polity Press in 2021.

Appendix 2 - Members of the Steering Committee

Ine Van Hoyweghen – coordinator / KVAB KMW / KU Leuven
Joos Vandewalle – KVAB KTW / KU Leuven
Marc De Mey – KVAB KMW / UGent
Lieven Verschaffel – KVAB KMW / KU Leuven
Johan Wagemans – KVAB KMW / KU Leuven
Luc Bonte – KVAB KNW
Luc Steels – KVAB KNW / VUB
Paul Verstraeten – KVAB KTW
Hugo De Man – KVAB KTW
Bart De Moor – KVAB KTW / KU Leuven
Anne-Mie Van Kerckhoven – KVAB KK / AMVK
Ann Dooms – Alumni JA / VUB
Inez Dua – KVAB Staff

RECENT POSITION PAPERS

61. Luc Bonte, Aimé Heene, Paul Verstraeten e.a. – *Verantwoordelijk omgaan met digitalisering. Een oproep naar overheden en bedrijfsleven, waar ook de burger toe kan/moet bijdragen*, KVAB/Klasse Technische Wetenschappen, 2018.
62. Jaak Billiet, Michaël Opgenhaffen, Bart Pattyn, Peter Van Aelst – *De strijd om de waarheid. Over nepnieuws en desinformatie in de digitale mediawereld*, KVAB/Klasse Menswetenschappen, 2018.
63. Christoffels Waelkens. – *De Vlaamse Wetenschapsagenda en interdisciplinariteit. Leren leven met interdisciplinaire problemen en oplossingen*, KVAB/Klasse Natuurwetenschappen, 2019.
64. Patrick Onghena – *Repliceerbaarheid in de empirische menswetenschappen*, KVAB/Klasse Menswetenschappen, 2020.
65. Mark Eyskens – *Als een virus de mensheid gijzelt. Oorzaken en gevolgen van de Coronacrisis*, KVAB/Klasse Menswetenschappen, 2020.
66. Jan Rabaey, Rinie van Est, Peter-Paul Verbeek, Joos Vandewalle - *Maatschappelijke waarden bij digitale innovatie: wie, wat en hoe?, KVAB - Denkersprogramma 2019*, KVAB/Klasse Technische Wetenschappen, 2020.
67. Oana Dima (auteur), Dirk Inzé, Hubert Bocken, Pere Puigdomènech, René Custers (eds)., *Genoombewerking voor veredeling van landbouwgewassen. Toepassingen van CRISPR-Cas9 en aanverwante technieken*, ALLEA-KVAB/Klasse Natuurwetenschappen, 2020.
68. Marie-Claire Foblets, *De multiculturele samenleving en de democratische rechtsstaat – Hoe vrijwaren we de sociale cohesie?*, KVAB/Klasse Mensweten-schappen 2020
69. Joost Van Roost, Luc Van Nuffel, Pieter Vingerhoets e.a., *De rol van gas in de Belgische energietransitie – Aardgas en Waterstof,* KVAB/Klasse Technische Wetenschappen, 2020.
70. Richard Bardgett, Joke Van Wensem, *Bodem als natuurlijk kapitaal* – KVAB Denkersrapport 2020, KVAB/Klasse Technische Wetenschappen, 2021
71. Jos Smits e.a., *Multifunctionele eilanden in de Noordzee*, KVAB/Klasse Technische Wetenschappen, 2021.
72. Elisabeth Monard, red., *Kunst, Wetenschap en Technologie in Symbiose*, KVAB/Klasse Technische Wetenschappen, 2021.
73. Jan Wouters, Maaike De Ridder, *De problematiek van de rechtsstaat en democratische legitimiteit binnen de Europese Unie*, KVAB/Klasse Menswetenschappen, 2021.
74. Hilde Heynen, Bart Verschaffel, e.a., *Architectuurkwaliteit vandaag, Reflecties over architectuur in Vlaanderen*, KVAB/Klasse Technische wetenschappen en Klasse Kunsten, 2021.
75. Godelieve Laureys & Kristiaan Versluys e.a., *Language Matters, Taalgebruik en taalbeleid aan de Vlaamse universiteiten,* KVAB/Klasse Menswetenschappen, 2022.
76. Bea Cantillon, *Het armoedevraagstuk en de tragiek van de welvaartsstaat, Zeven termen voor een nieuw sociaal contract*, KVAB/Klasse Menswetenschappen, 2022.
77. Joos Vandewalle, Marc Acheroy e.a *Een oproep tot een versnelde digitale transformatie voor België,* ARB/KVAB, 2022.
78. Jo Tollebeek, Marc Boone en Karel van Nieuwenhuyse, *Een Canon van Vlaanderen, Motieven en bezwaren*, KVAB Klasse Menswetenschappen, 2022.
79. Luc Taerwe e.a., *Duurzaam Beheer van Infrastructuur, Niet alleen een kwestie van budgetten*, KVAB/Klasse Technische Wetenschappen, 2022.
80. Willem Salet, Marleen Spiekman, Staf Roels, Tom Coppens, Ivo Van Vaerenbergh, *Naar klimaatneutrale woongebouwen in 2050, KVAB Denkersprogramma 2022*, KVAB/Klasse Technische Wetenschappen, 2022.
81. Sabina Leonelli, Stephan Lewandowsky, *De reproduceerbaarheid van het onderzoek in Vlaanderen: Feitenonderzoek en aanbevelingen – KVAB Denkersrapport 2022*, KVAB/Klasse Technische Wetenschappen en Menswetenchappen, 2022.
82. Elisabeth Monard e.a., *Vrij onderzoek noodzakelijk voor maatschappelijke uitdagingen, Ruimte voor wetenschap op initiatief van de onderzoeker*, KVAB/Klasse Technische Wetenschappen, Klasse Menswetenschappen, Jonge Academie, 2023.
83. Herman De Dijn, Gita Deneckere, Danny Praet, Jo Tollebeek, Sabine Verhulst, *Een noodzakelijk goed., Over het blijvend belang van de geesteswetenschappen*, KVAB/Klasse Menswetenschappen, 2023.