Open Letter: We are not ready for manipulative AI – urgent need for action

Voor de Nederlandstalige versie van deze open brief, klik hier.
Pour la version française de la lettre ouverte, cliquez ici.
A shortened version of this letter has been published on Euractiv on 3 April 2023.

By Nathalie A. Smuha, Mieke De Ketelaere, Mark Coeckelbergh, Pierre Dewitte and Yves Poullet - 31 March 2023

Chatbots and other human-imitating artificial intelligence (AI) applications are conquering an increasingly important place in our lives. The breakthrough of ChatGPT also marked the breakthrough of AI to the public, even if this technology has been around for decades. The possibilities raised by the latest AI developments are fascinating, but the fact that something is possible does not yet make it desirable.

Given the ethical, legal and social implications of AI, the question of its desirability – especially as regards its form, purpose, function, capabilities, requirements and safeguards – is becoming increasingly pressing. For example, we know that chatbots, like other AI systems, can contain biases and generate discriminatory output. They may also "hallucinate" or make a statement with great certainty that is completely disconnected from reality, as well as produce hateful, misinformation-based, or other problematic language. Their opaque operation and unpredictable evolution exacerbate this problem.

But one of the main risks associated with human-imitating AI was reflected in the recent chatbot-incited suicide in Belgium: the risk of manipulation. While this tragedy illustrates one of the most extreme consequences of this risk, emotional manipulation can also manifest itself in more subtle forms. As soon as people get the feeling that they interact with a subjective entity, they build a bond with this "interlocutor" – even unconsciously – that exposes them to this risk and can undermine their autonomy. This is hence not an isolated incident. Other users of text-generating AI also described its manipulative effects.

No understanding, nevertheless misleading

Companies that provide such systems easily hide behind the fact that they don’t know which texts their systems generate, and instead will point to their many advantages. Problematic consequences are dismissed as an anomaly that will soon improve, given that the technology is still evolving. Teething problems, which will be

solved with a few quick technical fixes.

After an American journalist reported last month how he had tested Microsoft's Bing AI Bot and was presented with the same text pattern as the Belgian victim (from love declarations to exhortations to leave his wife), Microsoft took some measures, like limiting the number of chats that could be exchanged. But similar bots exist on numerous websites, without any restrictions, and even Microsoft already started loosening these restrictions. In addition, numerous apps are also specifically aimed at providing chatbots with a 'personality', which further increases the risk of emotional manipulation.

Most users realize rationally that the bot they are chatting with is not a person and has no real understanding, but that it is merely an algorithm that predicts the most plausible combination of words based on sophisticated data analysis. It is, however, in our human nature to react emotionally to realistic interactions, even without wanting it. This also means that merely obliging companies to clearly indicate that “this is an AI system and not a human being” is not a sufficient solution.

Everyone is vulnerable

Some individuals, because of their age or mental state, are more susceptible than others to the effects associated with such realistic systems, and to their manipulative risks. For instance, the fact that children can easily interact with chatbots that first gain their trust and then not only spew hateful, conspiracy-oriented or pornographic language, but also encourage suicide, is alarming to many.

Others particularly susceptible are those without a strong social network, or those who are lonely or depressed – precisely the category which, according to the creators of the chatbots, can get the most 'use' from such systems. The fact that there is a loneliness pandemic, and that timely psychological help is lacking almost everywhere, only contributes to this issue. Yet it is important to underline that everyone can be susceptible to the effects of such realistic systems. After all, the emotional response they elicit happens automatically, even without us realizing it.

“Human beings, too, can generate problematic text, so what is the problem”, is a frequently heard response. But AI systems function on a much larger scale, so the damage they can cause is far greater too. And if it had been a human being communicating with the Belgian victim, we would have classified this as incitement to suicide and failure to help a person in need – offenses punishable by imprisonment.

“Move fast and break things”

How come these AI systems are available without any restrictions or required specifications? The call for regulation is often silenced by the fear that “regulation should certainly not stand in the way of innovation”. The Silicon Valley motto “move fast and break things” – meaning fast, experimental and disruptive innovation – reflects the idea that we should let AI inventors do their thing, for we have no idea yet of the marvelous benefits the technology can offer us.

The problem, however, is that the technology is capable of also literally breaking things – including human lives. This requires a more responsible approach, and a better balance between the precautionary principle and the innovation principle. Compare this with other technological developments. If a pharmaceutical company wants to market a new drug against disease X, it cannot simply claim that it does not know what the effect will be, but that it is certainly innovative and groundbreaking. The developer of a new car will also have to test the product extensively for all kinds of incidents and demonstrate that it is safe before the car can be marketed. Is it so far-fetched to expect the same from AI developers?

As entertaining as chatbots can be, they are more than just a toy, but can have very real consequences for the people who use them. The least we can expect from their developers is that they take up their responsibility, and only make these systems available when there are sufficient safeguards against harm. The creators and providers of chatbots should therefore not evade their moral and legal responsibility by stating that they have no idea how their system works and how it will react.

New rules: too little, too late

The European Union is currently working on new legislation that will impose stronger rules on “high-risk” AI systems and stricter liability on their suppliers, the so-called AI Act. However, the original proposal does not classify chatbots and generative AI systems as “high risk”, and their providers must only inform users that it is a chatbot and not a human being. A prohibition on manipulation was included, but only insofar as the manipulation leads to 'physical or mental harm', which is by no means easy to prove.

We can only hope that both member states and parliamentarians will strengthen the AI Act’s text during the ongoing negotiations and provide better protection against AI’s risks. A strong legislative framework will not stand in the way of innovation, but can actually encourage AI developers to innovate within a framework of values that we cherish in democratic societies. We can, however, not wait for the AI Act, which in the best case scenario will only come into effect in 2025. Given the great speed with which new systems are introduced, this legislative initiative already risks being too little, too late.

What now?

We therefore call to urgently set up awareness campaigns that better inform people of the risks associated with AI systems, and that encourages AI developers to take their responsibility. A shift in mindset is needed to ensure the risks of AI are first identified, tested and tackled, before the latest application is made available. Education has an important role to play here, at all levels. Yet there is also an urgent need to invest more into research on AI’s impact on fundamental rights, including the right to physical and moral integrity. Finally, we call for a wider public debate about the role we wish to give AI in society, not only in the short term but also in the longer term.

Let us be clear: we too are fascinated by the capabilities AI systems bring. But that does not prevent us from also wanting to ensure that these systems are human rights-compliant. The responsibility for this lies not only with AI developers and providers, but also with our governments at national, European and international level. They

should adopt a protective legal framework with strong safeguards and prior verifications as soon as possible. This also requires expert (consultation) bodies that anticipate the risks in a multidisciplinary and multistakeholder manner.

In the meantime, we ask that all necessary measures be taken – through data protection law, consumer law, and if need be the imposition of targeted moratoria – to prevent the tragic case of our compatriot from repeating itself. Let this be a wake-up call to us all. The AI playtime is over: it's time to draw lessons and take responsibility.

Authors:

- Nathalie A. Smuha, legal scholar & philosopher, KU Leuven
- Mieke De Ketelaere, engineer, Vlerick Business School
- Mark Coeckelbergh, philosopher, University of Vienna
- Pierre Dewitte, legal scholar, KU Leuven
- Yves Poullet, legal scholar, University of Namur

Co-signatories:

- Adriana Giunta, legal scholar & consultant, University of London
- Aimen Taimur, lawyer, Tilburg University
- Alberto Delgado, engineer, Universidad Nacional de Colombia
- Alessio Rocchi, pedagogist, Istituto Universitario Salesiano Torino Rebaudengo
- Alexander Antonov, doctoral candidate, TalTech University
- Alberto Villa, engineer
- Aleksandra Kuczerawy, legal scholar, KU Leuven
- Anca Radu, legal scholar, European University Institute
- Andoni Iturbe Mach, jurist, Basque Parliament
- Ann-Katrien Oimann, philosopher & legal scholar, Royal Military School & KU Leuven
- Anne-Lise Sibony, legal scholar, UCLouvain/KULeuven
- Anne-Mieke Vandamme, virologist/bio-computerscientist, KU Leuven
- Annette Kleinfeld, philosopher, University of Applied Sciences (HTWG Konstanz), Germany
- Annick De Paepe, psychologist, UGent
- Antoinette Rouvroy, legal scholar and philosopher, UNamur
- Anton Vedder, philosopher, KU Leuven
- Aron Turner, BigMother.AI
- A Duroisin, honorary auditor, IRE
- Bart Preneel, engineer, KU Leuven
- Benoit Macq, engineer, UCLouvain
- Bert Peeters, legal scholar, KU Leuven
- Bieke Zaman, human-Computer Interaction / media studies, KU Leuven
- Bilel Benbouzid, professor of sociology, Université Gustave Eiffel, Paris
- Birgit Schippers, legal scholar, University of Strathclyde

- Boris Thienert, IT engineer
- Brett Patching, behaviour designer, The Behaviour Bureau
- Carl Mörch, psychologist, FARI / Université Libre de Bruxelles
- Carlos Andrés Salazar Martínez, engineer & philosopher, Universidad EAFIT
- Catherine Jasserand, legal scholar, KU Leuven
- Catherine Van de Heyning, legal scholar, Universiteit Antwerpen
- Cécile de Terwangne, legal scholar, Université de Namur
- Charalambos Tsekeris, sociologist, Greek National Centre for Social Research
- Charlotte Ducuing, legal scholar, KU Leuven
- Christine Gealy
- Claire Boine, AI lawyer, University of Ottawa
- Daniel Soto Zeevaert, data Scientist, TIMi
- Daniela Spajic, legal scholar, KU Leuven
- David Doat, philosopher, Catholic University of Lille
- David Geerts, social scientist, KU Leuven Digital Society Institute
- Debbie Esmans, meemoo vzw
- Debra Rowe, content creator
- Dennis Wilson, AI and data science, University of Toulouse
- Diego Calvanese, computer engineer, Free University of Bozen-Bolzano
- Dietrich Heiser, Consultant, i-LUDUS bv
- Dirk Remans, IT
- Eirini Christinaki, computerscientist, KU Leuven
- Elfi Baillien, work and organization studies, KU Leuven
- Elise Degrave, legal scholar, UNamur
- Eva Thelisson, lawyer, AI Transparency Institute
- Federico Cabitza, IT engineer, University of Milano-Bicocca
- Firouzeh Rosa Taghikhah, business analytics, University of Sydney
- Francesco Marzoni, engineer
- Francis Wyffels, engineer (robotica & AI), UGent
- Frank Maet, philosopher, LUCA / KU Leuven
- Frankie Schram, law & public administration, KU Leuven
- Frederic Heymans, communication scientist, Kenniscentrum Data & Maatschappij
- Friso Bostoen, legal scholar, European University Institute / KU Leuven
- Gabriela Cucos, visual artist
- Gaëlle Fruy, legal scholar, Université Saint-Louis - Bruxelles
- Gaëlle Vanhoffelen, researcher, KU Leuven
- Gary Marcus, AI researcher, New York University (Emeritus)
- Geert Crombez, psychologist, UGent
- Geert De Deyn, engineer, Gedipro consulting
- Geert van Calster, legal scholar, KU Leuven; King's College London; Monash University

- Geertrui Van Overwalle, legal scholar, KU Leuven
- Geneviève Vanderstichele, legal scholar, University of Oxford/Raadsheer hof v beroep Gent, gedetacheerd naar het Digital Transformation Office
- Giacomo Franco, teacher & project manager, Ministry of Education, Italy
- Gianluca Bontempi, machine learning, Université Libre de Bruxelles (ULB)
- Giovanni Barbuti, consultant
- Griet Verbeke, House of AI, VIVES/Kulak
- Guido Boella, computer science and philosophy, University of Torino / Italian Society for the Ethics of AI (SIpEIA)
- Hannah Carlota Osaer, jurist
- Hannes Cools, communication scientist, AI, Media, and Democracy Lab, Universiteit van Amsterdam
- Hans Lombaert, criminologist, RU Gent
- Hans Radde, philosopher, Vrije Universiteit Amsterdam
- Heidi Mertes, medical ethics, Universiteit Gent
- Hendrik Blockeel, engineer, KU Leuven
- Igor Barshteyn, information security & data privacy
- Ihor Gowda, writer
- Iñaki Vicuña de Nicolás, jurist, CENDOJ / CGPJ (Emeritus)
- Ine Van Hoyweghen, sociologist, KU Leuven
- Isaac Crivillés i García, engineer, CivicAI.cat
- Isabel Barberá, privacy engineer / juriste, Rhite
- Ivo Poje, Humannet
- James Gealy, electrical engineer
- James Harland, computer scientist, RMIT University
- Jan Hauters, postgraduate researcher, UCL-IOE-CCM-Knowledge Lab, London | 4Dclass Education Technology, Beijing
- Jan Kleijssen, legal scholar & consultant, LUISS University Rome
- Javier Gállego Diéguez, public health
- Jean-Jacques Quisquater, engineer, UCLouvain
- Jochen De Weerdt, business and process analytics, KU Leuven
- Johan Decruyenaere, medicine, UGent
- Jonathan Resnick, software engineer
- Joost Vennekens, computer scientist, KU Leuven
- Jose María Zavala-Pérez, social impact researcher & consultant
- Joshua C. Gellers, political scientist, University of North Florida
- Joshua J. Morley, emerging technology specialist, Morley Technologies
- Jozefien Vanherpe, bestuurskundige, KU Leuven
- Juan Gérvas, physician, Equipo CESCA
- Juan Manuel Velasquez, mathematician
- Karianne J. E. Boer, criminologist and sociologist of law, Vrije Universiteit Brussel
- Katherine Pardo, lawyer
- Kathryn Conrad, english professor, University of Kansas
- Katrien Verbert, computerscientist, KU Leuven

- Kiril Aleksovski, data engineer
- Klaus Speidel, philosopher, University of Applied Arts Vienna
- Kristof Hoorelbeke, clinical psychologist, UGent
- Kyle Shaffer, NLP Researcher
- Laura Drechsler, legal scholar, KU Leuven/Open Universiteit
- Laurens Naudts, legal scholar, AI, Media and Democracy Lab – Universiteit van Amsterdam
- Laurent Hublet, entrepreneur & philosopher, ULB / Solvay Brussels School
- Leen d'Haenens, social scientist, KU Leuven
- Lieven De Lathauwer, engineer, KU Leuven
- Lino Helbling, career counselor
- Lode Lauwaert, philosopher, KU Leuven
- Luis Caires, computer scientist, NOVA University Lisbon
- Maarten Buyl, AI researcher, UGent
- Maciej Majewski, physicist, professor emeritus
- Magali Legast, engineer, UCLouvain
- Maite Sanz de Galdeano, lawyer, OdiseIA
- Malcolm Muckle, amateur philosopher
- Manuela Battaglini, digital ethicist & lawyer, Transparent Internet
- Marc Rotenberg, law and technology, Center for AI and Digital Policy
- Margot van der Goot, communication scholar, University of Amsterdam
- Maria Vieira, student, Universidade do Porto
- Marian Verhelst, engineer, KU Leuven en Imec
- Mark Brakel, policy, Future of Life Institute
- Mark Depauw, philologist, KU Leuven
- Marlon Domingus, data protection, Erasmus University Rotterdam
- Martin Meganck, engineer/ethicist, KU Leuven
- Massimiliano Simons, philosopher, Maastricht University
- Maximilian Rossmann, technology assessment, Maastricht University
- Mercedes Pérez-Fernández, physician, Equipo CESCA
- Merve Hickok, AI ethicist, Center for AI and Digital Policy
- Michel Herquet, physicist, B12 Consulting
- Michiel De Proost, philosopher, Universiteit Gent
- Moa Mörner, practical philosophy / data protection officer
- Michel Schellekens, computer scientist, University College Cork
- Nathanaël Ackerman, engineer, AI4Belgium SPF BOSA
- Nele Roekens, legal officer, Unia
- Nick Brown, software engineer, Google
- Nick von Beroldingen, software engineer
- Nico Miailhe, STS, The Future Society (TFS)
- Nikos Koutras, legal scholar, Curtin University

- Norberto Patrignani, computer ethicist, Politecnico of Torino
- Oliver Bown, creative AI, University of New South Wales
- Orian Dheu, legal scholar, KU Leuven
- Ozturk Taspinar, innovation consultant, KPMG
- Pablo Martínez Ramil, lawyer, UPOL/Taltech
- Paolo Petta, engineer, Universität Wien
- Patryk Ciurak, legal scholar, University of Gdańsk
- Paul De Hert, legal scholar, Vrije Universiteit Brussels
- Peggy Valcke, legal scholar, KU Leuven
- Plixavra Vogiatzoglou, legal scholar, KU Leuven
- Rachel Alexander, CEO, Omina Technologies
- Ralf De Wolf, media studies, Universiteit Gent
- Rene Walter, journalist, Good Internet
- Rieks op den Akker, mathematical engineer, Universiteit Twente
- Risto Uuk, policy researcher, Future of Life Institute
- Robin Schrijvers, researcher AI & data, Hogeschool PXL
- Roger Vergauwen, philosopher, KU Leuven (Emeritus)
- Rosamunde Van Brakel, criminologist, Vrije Universiteit Brussel
- Rosanna Baltzer, data scientist & cultural theorist
- Rupert Russell, learning technologist
- Sally Wyatt, science & technology studies, Maastricht University
- Samuel Ramírez, engineer, National Polytechnic Institute
- Saskia Dörr, digital reponsible management, WiseWay
- Scott L. Burson, computer scientist
- Seppe Segers, philosopher, Universiteit Gent & Universiteit Maastricht
- Siegfried Nijssen, computer scientist, UCLouvain
- Sigrid Sterckx, ethicist, UGent
- Società Italiana per l'Etica dell'Intelligenza Artificiale, Philosophy / Computer Science / Law, SIpEIA
- Sofia Palmieri, legal scholar, UGent
- Solomon Williams, accountant, Stipenda
- Sophie Stalla-Bourdillon, legal researcher, VUB & Immuta
- Srivathsan Karanai Margan, insurance consultant, Tata Consultancy Services
- Stefan Ramaekers, pedagogue and philosopher, KU Leuven
- Stefanoi Moi, digital innovator, The Value Hub
- Stephanie Rossello, legal scholar, KU Leuven
- Tanya de Villiers-Botha, philosopher, Stellenbosch University
- Teodora Lalova-Spinks, legal scholar, KU Leuven
- Thierry Léonard, legal scholar, Université Saint-Louis - Bruxelles
- Thomas Buule, developer
- Thomas Gils, legal scholar, Kenniscentrum Data & Maatschappij

- Thomas Hildebrandt, computer scientist, Copenhagen University
- Tianxing Dwight Xia, history & philosophy, University of Sydney
- Tias Guns, computer scientist, KU Leuven
- Tijl De Bie, data scientist / AI researcher, UGent
- Tim Christiaens, philosopher, Tilburg University
- Tinne De Laet, engineer, KU Leuven
- Toby Walsh, artificial intelligence, AI Institute UNSW Sydney
- Tomas Folens, ethicist, KULeuven/VIVES
- Tsjalling Swierstra, philosopher, Maastricht University
- Veronika Romhány, multimedia artist, KU Leuven - LUCA School of Arts
- Victoria Hendrickx, legal scholar, KU Leuven
- Vincent Vandeghinste, language technologist, KU Leuven
- Vinciane Gillet, lawyer, GILLET-LEX
- Warren Bell, physician, University of British Columbia
- Weina Jin, AI researcher, Simon Fraser University
- Wim Van Biesen, doctor, UGent
- Wolfgang Keck, senior advisor, Digital Society
- Wouter Baetens
- Yves Persoons, communication manager, KU Leuven

This letter solely reflects the views of the authors and co-signatories, and does not represent the position of the Faculty or the University.

Back to Blog

Last update: 14 Jun 2023