AI at the Battlefield of Human Mind

Inaugural address, delivered by
Prof. dr. Marie Šafář Postma

Tilburg University

Understanding Society

Marie Šafář Postma is Professor of Computational Cognitive Science at Tilburg University in the Netherlands. After obtaining a master's degree in Theoretical Linguistics at the University of Tromsø (1999), she graduated cum laude in Logic at the University of Amsterdam (2001). In 2006, she obtained her PhD from the Institute of Logic, Language and Computation at the University of Amsterdam with a dissertation focusing on modal-logic models of speech patterns, based on a theory anchored in biological mechanisms of sound frequency perception. In her research she makes use of AI methods and techniques to model human cognitive processes, with applications in the domain of education, health and safety & security.

She is the principle investigator of a number of collaborative research projects in the domain of artificial intelligence, most notably the NWO funded project STEADFAST exploring the development of semi-autonomous human-drone swarm teams for rapid response in military and civil contexts. While at Tilburg University, Marie created several new interdisciplinary educational programs in data science, cognitive science and artificial intelligence; in 2023, she initiated the development of a new master's program in cybersecurity and artificial intelligence. The success of these educational initiatives led to the establishment of a new Department Cognitive Science and Artificial Intelligence for which Marie served as Head of Department in 2021-2024.

AI at the Battlefield of Human Mind

Prof. dr. Marie Šafář Postma

Inaugural lecture,
Delivered in adapted form on the occasion of the public acceptance of the appointment of Professor in Computational Cognitive Science at Tilburg School of Humanities and Digital Sciences on August, 30th, 2024, by Marie Šafář Postma.

ISBN: 978-94-6167-526-2

www.tilburguniversity.edu

AI at the Battlefield of Human Mind

1. Introduction

Buchstaben zu empfinden, sie nicht nur mit den Augen in Büchern zu lesen, - einen Dolmetsch in mir selbst aufzustellen, der mir übersetzt, was die Instinkte ohne Worte raunen, darin muß der Schlüssel liegen, sich mit dem eigenen Innern durch klare Sprache zu verständigen, begriff ich.

- Gustav Meyrink, Der Golem (1915)

Dear Rector Magnificus,

Dear highly esteemed colleagues, friends, and family,

It is a great honor to be standing in front of you today.

The title of my inauguration lecture – "AI at the battlefield of human mind' – was deliberately chosen to be ambiguous. It can be interpreted in three different ways.

Under the first interpretation, the title refers to the fact that artificial intelligence has acquired capabilities that we thought were exclusive to the human brain. It cannot be denied that AI is quite successfully competing with – battling – the power of individual and collective human intelligence in many areas. Just ask chatGPT! Does that mean that in a not too distant future, AI may come to fully resemble human beings and develop a form of sentience or consciousness? To be able to answer this question, we need to understand what it means to be conscious. Despite all of us being experts on this deeply subjective experience, it turns out to be quite a challenge to translate it into scientific terms. How do I know what it is like to be 'you' or interpret 'you as being like me'? And if "for a conscious organism, there is something it is like to be that organism" (Nagel, 1974), how does this experience pertain to current AI capabilities? Contrary, perhaps, to our intuition, having a language is not a prerequisite. Even very simple worms like the soil-dwelling *Caenorhabditis elegans* have "elegant minds" that provide them with amazing sensory, social, and learning capacities (Ardiel & Rankin, 2010). Many cognitive scientists would argue that the fundamental building blocks of consciousness are our ability to perceive the external and internal world and to act on this

information so that the wonderful assembly of cells that is you or me can survive. In this view, consciousness is a mechanism that evolved to support homeostasis - the process by which biological systems maintain stable values of their vital parameters (such as temperature, glucose levels, blood pressure or water balance). If we consider operational parameters of AI systems (e.g., CPU temperature or memory usage) to be AI's internal states, provide them with external sensors and include self-monitoring and maintenance features in their design, would AI become conscious? Under these circumstances, would it still be fundamentally different from human beings or is it just an idle anthropocentric hope?

The second interpretation of the title 'AI at the battlefield of human mind' refers to the idea that increasingly, AI is enhancing humans. Different cognitive processes are competing for resources in a world that demands constant attention. Teaming up with AI brings about results that are superior to those human cognition can accomplish on its own. For example, in radiology, the highest accuracy is achieved when the interpretation of medical images is done by AI systems trained to detect subtleties that escape the human visual cognition, in collaboration with human experts who filter out false positives - as demonstrated by my colleague Sharon Ong and her team. AI is also very useful in dealing with cognitive overload. Let us consider the fact that external sensory information collected by the human perceptual system represents about 11 million bits per second. This information is presented to the brain for processing, but our conscious mind is unable to deal with such a data deluge. According to Encyclopedia Britannica, it can actually process only about 50 bits per second. What happens to the remaining 10 999 950 bits, you may ask? Our cognitive system appears to be designed in such a way that it can perform an enormous amount of data compression and filtering without endangering our chances of survival. As a result, though, human perception becomes a 'con job', a 'controlled hallucination', rather than a direct reflection of reality. The brain constructs our perceptual experience by combining predictions (based on past experiences and context) with only limited sensory input needed for corrections. Thanks to this controlled process, you are likely to understand what I am singing despite me not articulating every word clearly and some of you sitting in the back of the room. You have a pretty good impression of this stage and

the aula even though you are not constantly monitoring every detail of the environment and constructing an image of it from scratch. It also likely means that you just missed the fact that I used the verb "singing" instead of "saying" to refer to the activity that I am currently engaged in. Not a big deal unless life-or-death decisions have to be made depending on our perceptual precision. For such a scenario, let us imagine that the year is not 2024 but 1953, it is February 1, and large areas of the Zeeland, South Holland and North Brabant provinces have been flooded. Many people are sheltering on rooftops and inside houses, and there are farm animals in danger. Some are injured and require urgent medical help. At the time, the Netherlands had only one helicopter at its disposal. Now imagine that you are piloting this helicopter. It is dark and so you have a hard time detecting all the living beings waiting to be rescued. Within fractions of seconds, you need to make decisions about where first responders should be sent. Your cognitive system is fighting a battle about how to distribute your scarce attentional resources and you are growing tired. What if, though, you would have hundreds of drones equipped with computer vision and infrared and acoustic sensors that have been trained to scan a large area and detect living animals semi-autonomously, only requiring a quick validation from a human operator. Wouldn't it be of great value to you when deciding where to send help? However, in such instances of human-AI teaming, who is actually responsible, morally and legally? Is it you, the operator, the human-AI team as a whole, the AI system, the engineer who trained it or the government that decided to use this technology?

Finally, the third interpretation of the title of this lecture refers to the fact that the human mind is the grounds where wars are being waged using AI to capture our attention and influence our beliefs and behaviors. We tend to think of inter-state conflicts as involving physical weapons, but modern wars are primarily a social phenomenon, with battles taking place on search engines and social media. As I am sure you are aware, in the last 10-15 years, the population in all European countries has been exposed to these systematic hybrid war efforts that destabilize our democratic structures. Advancements in the field of AI and cognitive neuroscience brought about new possibilities of influencing the human brain with the help of so-called cognitive warfare. "Cognitive warfare is a strategy that focuses on altering how a target population thinks - and through that how

it acts" (Burda, 2023). By collecting and analyzing data about our online behavior and preferences, AI algorithms have come to know us better then we may know ourselves. Cleverly tailored messages, real and fake, have the ability to appeal to our emotions and capture our attention, as we have clearly experienced during the COVID-19 pandemic. How can we get AI on our team, to help us sharpen our senses and detect instances of disinformation, to protect us against cognitive manipulation? Does AI have a moral obligation to be truthful?

To be able to answer all these questions requires competence in the field of computational cognitive science - but also anthropology, political sciences, philosophy and communication studies. In this respect, the School of Humanities and Digital Sciences, where my research activities are situated, offers a unique opportunity to study the technology that has the power to influence the safety of humankind and of our democracy for better and for worse. In my lecture today, I will address the three perspectives on AI and human mind in due order from the point of view of my own research. Before we embark on this journey together, let me make a personal note. Artificial intelligence is currently a hot topic with many people perceiving it as both a tremendous opportunity and a threat. As an individual with an expiration date, I see the main value of AI research in that it offers us a deeper understanding of who we are, beyond the terms served to us by common sense. It comes with risks but at its best, AI technology offers a spiritual experience to us: Metaphorically, because by holding a mirror to human cognition, it provides us with the possibility to commune with what is otherwise inaccessible. And literally, because we can use it to achieve altered states of consciousness, as discussed in the second part of my talk.

2. The discipline of computational cognitive science

Demain est moins à decouvrir qu'à inventer.

-Gaston Berger, Phénoménologie du temps et prospective (1964)

From its start in the 1950s, the discipline of AI had as its aim to create machines capable of very human tasks such as natural language processing and translation, problem solving and decision making. One of the first steps to achieve such machine intelligence was the construction of a simulated neural network based on the principles of Hebbian learning. This principle is often summarized as 'neurons that fire together wire together'. For algorithms, this means that they adjust the weights of connections based on the correlation of artificial neuron activations, mimicking biological processes. Collaboration between cognitive science and AI has proven to be mutually beneficial with AI profiting from our knowledge about the brain (such as the principles of Hebbian learning), and neuroscience making use of AI modelling techniques to gain deeper understanding of neural structures and functions and their impact on human health and wellbeing. For example, within the Zero Poverty Project at Tilburg University, led by Margriet Sitskoorn, PhD researcher Valentina Sanchez Melchor, under the supervision of Çiçek Güven, Gonzalo Nápoles and me, explores graph-based data augmentation techniques to improve performance of machine learning models on functional magnetic resonance (fMRI) data, with the ultimate purpose to understand systematic relations between brain structures and growing up in poverty.

Computational models of cognition have also become the driving force behind our understanding of pathological brain functions (Read Montague et al., 2012), using, for example, brain measurements obtained with electro-encephalography (EEG) as input data (Hollander et al., 2016). This is a line of research pursued by my colleagues within the Computational Psychiatry lab, Marijn van Wingerden, Mosi Kia and the PhD researcher Zaman Nasrabadi, in whose project I am fortunate to be involved. EEG was invented in 1920 by the psychiatrist Hans Berger who placed electrodes on the human scalp and was able to record patterns of electrical activity and identify different brain wave patterns. Berger was motivated by a personal experience from his youth when during a military exercise, he fell off his horse and nearly died. Eerily, that same day, his sister, who was far away, got worried about Hans and asked their father to send a telegram asking if he was doing well. Berger considered this an example of 'spontaneous

telepathy' and set out to explore the mechanism of how thoughts could travel.[1] Berger himself later concluded that the electrical activity he recorded was not able to carry so far to explain psychic transference. It has, however, proven very useful to provide us with information about latent cognitive operations such as attentional focus, memory retrieval, and sensory processing. And interestingly, simultaneous recording of cerebral activity from two or more people socially engaged with one another, for example mothers interacting with their infants, reveals occurrences of inter-brain synchronization (so Berger's idea was not as outlandish as some people thought).

Computational cognitive models built on neurophysiological and behavioral data have proved immensely useful for all kinds of applications in the domain of education, safety and security, and health. Various research teams, including ours, have been able to obtain empirical evidence for processes that were previously mainly described in phenomenological terms, including spontaneous cognition in the default mode network (Lee, 2013), levels of consciousness (Seth & Bayne, 2022), and the interplay between attentional focus on external (sensory) and internal (interoceptive) perceptual events, often observed during mind wandering episodes (D'Mello, Tay, & Southwell, 2022; Garfinkel, Schulz, & Tsakiris, 2022). This information can be used in human-AI interaction and teaming to offer an assessment of the cognitive performance of the human user. For AI-based applications to be able to provide information and advice safely and effectively in high-stake situations, it is important that they take into consideration human limitations, including resource limitations, biases and heuristics (Tversky & Kahneman, 1974; Awad, et al., 2018; Hawkins, et al., 2019).

[1] https://www.sciencenews.org/article/hans-berger-telepathy-neuroscience-brain-eeg

3. Perception and attention as building stones of consciousness

What a bizarre animal we are that the only question we can ask in relation to our place in nature is "Mirror, mirror on the wall, who is the smartest of them all?"

- Frans de Waal, Are We Smart Enough to Know How Smart Animals Are?

Opinions differ on what the biggest achievements of human intelligence are. For some, it is the fact that "they put a man on the Moon" (REM, 1992), for others, it would be Richard Wagner's The Ring Cycle. Compared to that, the top achievements of AI as of today appear to be relatively modest. For example, it was considered an enormous AI breakthrough when AlphaGo was able to defeat a world champion in the game Go. Of course, this is a game that can be played even by children at a competitive level. Another often mentioned example is the language-transformer model DALLE-E, developed and trained by OpenAI. DALLE-E can create images from text captions, no matter how abstract or unreal these descriptions are. The model was trained on more than 250 million pairs of texts and images and holds about 12 billion parameters and its outcomes are certainly impressive. However, despite being very intelligent, if you ask DALLE-E to produce an image of a scientist, it comes up with a simple drawing of a person wearing glasses and a lab coat with a pocket full of pens. This image comes very close to the stereotypical description of a scientist from 1975, who would be "a man who wears a white coat and works in a laboratory. He is elderly or middle aged and wears glasses...he may wear a beard...he is surrounded by equipment: test tubes, Bunsen burners, flasks and bottles...he has to keep dangerous secrets" (Chambers, 1983). This is very different to people that I consider to be inspiring researchers, such as the primatologist Frans de Waal, the cognitive neuroscientists Antonio Damasio and Anil Seth, or the founder of affective computing computer scientist Rosalind Picard. Does it mean that at their current best, generative AI models can only reproduce human stereotypes and biases? That would be an unfair conclusion given the many instances of AI-produced artistic achievements. These include the completion of Beethoven's last symphony or the captivating art project of Jake Oleson 'Given Again', depicting an artist that seems to transform into the roots of a tree, expressing Oleson's state of mind upon discovering that his father was not his biological parent. As for Go, the 37th move taken by AlphaGo in Game Two was described by commentators as being absolutely unexpected, strange, and beautiful. It took Lee Sedol, the human opponent, fifteen minutes to formulate a response only to be beaten in the game.

Let us grant current or future AI systems creative skills next to exceptional information processing and decision-making abilities. Still, one could argue, AI could make the right decision at the right time without having a sense of what it is like to be itself, without having a subjective experience of that moment (Seth, 2024). How many of you subscribe to this human exceptionalism idea? Perhaps you have heard of the software engineer Blake Lemoine who was fired by Google in 2022 for violating employment and data security policies.[2] Lemoine became persuaded that the AI system he was interacting with gained sentience or consciousness because it felt anxious when confronted with certain conversation topics and even provided advice on which religion to convert to.[3] Was Lemoine right?

There are two perspectives one can take when answering this question, which roughly correspond to what is known as the hard and the soft problem of consciousness. According to the first perspective, there is no way to determine the answer. Using the philosophical zombie argument, popularized by David Chalmers, it is possible that I am the only conscious organism on this planet and all others, including you in the audience here today, act like me but in fact completely lack any subjective point of view. Put differently, "human beings never directly perceive other minds; they infer them" (Harris, 2024). Lemoine was simply extending this inference to the AI he interacted with, similarly to the roughly 4.000 men in Japan who 'married' a hologram designed by the company Gatebox as the perfect wife.[4] The idea that the inference of sentience in others goes hand in hand with our feelings for them is perfectly captured in the British children's book The Velveteen Rabbit from 1922: "Real isn't how you are made, it's a thing that happens to you when someone loves you" (Margery Williams, The Velveteen Rabbit). By virtue of this reasoning, whether or not AI is conscious is in the eye of the beholder.

The second perspective on consciousness is a scientific one. For determination of consciousness is not just a philosophical exercise; it carries substantial ethical and medical implications, for example, when ascribing moral responsibility

[2] https://www.annualreviews.org/content/journals/10.1146/annurev-psych-013123-123421
[3] https://www.newsweek.com/google-ai-blake-lemoine-bing-chatbot-sentient-1783340
[4] https://www.cbc.ca/documentaries/the-nature-of-things/i-love-her-and-see-her-as-a-real-woman-meet-a-man-who-married-an-artificial-intelligence-hologram-1.6253767

or assessing a patient with a brain injury (Melloni et al., 2021). Though there are currently at least 22 scientific definitions of consciousness and many ideas how to test for it (Bayne, et al., 2024), there is a strong case to be made for the role of consciousness as a mechanism that evolved to support survival (Damasio & Damasio, 2024). This mechanism makes use of sensory stimuli that originate both within and outside of the organism, so called interoception and exteroception. "The self is a repeatedly reconstructed biological state" (Damasio, 1994). The sensory information is bound into a conscious experience that is largely determined by what proved to be relevant to the organism's survival in the past - 'remembering the past and imagining the future'. This is what we experience when we are mind wandering, a task-unrelated mental activity we appear to be engaged in about 30-50% of our waking moments (Smallwood & Schooler, 2015). My colleagues Myrthe Faber, Mariana Dias da Silva and myself explored how when the mind wanders, it focuses on the cognitive structures that create, regulate, and assess behavior to support the organism's homeostasis, including reasoning about our own beliefs (so-called metacognition). This process is measurable in the electrical activity of the brain, as we confirmed in a study of mind wandering conducted with EEG (Dias da Silva, Gonçalves, Branco & Postma, 2022).

I cannot think of any compelling arguments why artificial intelligence could not be equipped with interoceptive and exteroceptive mechanisms and the overarching goal of maintaining its own mechanical homeostasis. After all, just like Large Language Models (LLMs), our perception, too, is prominently anchored in predictions about what is most likely to happen next. An example of such predictions comes from the perception of complex sounds. As demonstrated in experiments with the so-called missing fundamental (Schneider et al., 2005) which we conducted some years ago (Postma-Nilsenová & Postma, 2013; Tsoumani & Postma-Nilsenová, 2013), listeners perceive the pitch of a sound that corresponds to the fundamental frequency, even when that fundamental frequency is physically absent in the sound wave. This is because the human brain supplements the missing information based on the harmonic series. For example, if the sound consists of the harmonics at 200 Hz, 300 Hz and 400 Hz, based on past experiences, the auditory perceptual system would infer that the pitch is 100 Hz, whether or not it is present. This can be quite useful because traditional telephone systems transmit audio signals within a limited frequency range (ca. 300 - 3400 Hz), which excludes the fundamental frequency of a

person's voice. However, we can still perceive if the speaker is asking a question or making a statement by inferring the pitch correctly from the harmonics that are transmitted. Perceptual decisions remembered from the past thus determine our hallucinations. That is why musicians who play instruments that are rich in timbre (e.g., violoncello or saxophone) infer the pitch in stimuli with a missing fundamental differently than musicians who play clearly pitched instruments (e.g., piano or guitar).

There are, however, moments in the human experience when the predictive perceptual mechanisms break down. Think of the vast endless space around you when standing on top of a mountain, when feeling weightless on a roller-coaster in the Efteling or when hallucinogens lead you to believe that you are flying. In such moments, our sensory system is unable to provide crucial information about the position and orientation of the body and its parts in space. The lack of useful referents brings our spatial awareness and sense of balance to a halt. It creates the need for a bottom-up driven sensory processing that is fully anchored in the moment. This can be both scary and wonderful at the same time. The sensation associated with such an event, *awe*, lies at the core of self-transcendent experiences (Yaden et al., 2017). AI technology provides us with means to achieve such states in virtual reality, as in the case of the simulated Overview Effect (watching the Earth from space) or the Is-ness multi-person distributed VR environment where participants lose the feeling of their self-other bodily boundary (Glowacki et al., 2022). Once the hallucination of who we are in relation to others breaks down, we feel small and insignificant, but also more connected, as demonstrated in studies of PhD researcher Anna van Limpt-Broers who is supervised by Max Louwerse and me (Graziosi & Yaden, 2021; van Limpt-Broers, Postma, & Louwerse, 2024; van Limpt-Broers, et al., 2024). Blurring the boundaries of self opens us to new beliefs and to learning (Broers, Louwerse, & Postma, 2020). It seems highly unlikely that AI itself could ever appreciate as transformative when its predictive powers come to a grind. Somewhat ironically, then, what might be distinguishing us from AI is not the experience of consciousness but rather the ability to lose our-*selves*.

4. Human-AI teaming

To the electrodes responsive to brain activities associated with thought processes, muscular movements and emotional states, may be added electrical sensing elements responsive to corresponding body reactions, such as movement, pulse and breathing rates, temperature, blood pressure, skin resistance, and blood composition. Once sufficient transfer of information from man to electrical circuit is achieved, machines can be built which will 'understand' that information with speed and accuracy possible only with machines. The machine-to-man coupling area of 2012 A.D. is quite unknown to today's civilisation. The information which a machine can obtain and store from a person in a few minutes will exceed the fruits of a lifetime of man-to-man communication.

- R.M.Page, Man-Machine Coupling - 2012 A.D. (1962)

There are many examples of contexts where individuals feel a strong sense of unity with a group, the boundaries of their physical and mental self softening and becoming indistinct. These include religious rituals, sport events, musical festivals, and political protests, but also some occupations. First responders, for example, working together in high-stress situations, develop a collective agency when thinking, feeling and acting synchronously (Shteynberg, et al., 2022; Shteynberg, et al., 2023). The emergence of group cognition is supported by behavioral imitation and adaptation. We have been able to confirm empirically that people tend to mimic each other's speech patterns (Postma-Nilsenová, Brunninkhuis, & Postma, 2013; Mattheij, Postma, & Postma, 2015; Rombout & Postma, 2022), gestures, and even develop synchronized cerebral activity, as in the already mentioned mother-infant dyads. Imitation of nonverbal behavior helps group members understand each other's intentions and actions without verbal cues, it fosters empathy and rapport and enhances teamwork. It builds trust and solidarity within a group. When group members imitate each other, it signals mutual respect and willingness to cooperate. These are all very human mechanisms. How can we achieve a human-machine coupling? Can a collective mind emerge in teams consisting of humans and machines?

Improving human-AI collaboration and teaming has been on the agenda of AI researchers for more than 60 years, as illustrated by the quote from Robert Morris Page, the director of research for the U.S. Naval Research Laboratory. It was primarily this effort that led to the establishment of the field of affective

computing - the development of technological solutions to measure, identify, and process human emotions (Picard, 1997). When in the 1990s the computer scientist Rosalind Picard argued that computers should be able to interpret human emotional states and adapt to them, she was advised not to share this opinion at scientific conferences because it could be damaging to her career. After all, computers were supposed to be rational machines and having emotions could interfere with their performance. "Who really needs an emotional toaster?" (Breazeal & Picard, 2006). I am not sure about the toaster but nowadays, we do not doubt any more that computers need emotional intelligence to understand human decision-making and to be able to collaborate with us.

Let me mention a few examples. As demonstrated by my colleagues Bosong Ding, Murat Kirtay and Giacomo Spigler, when a robot imitates human head movements in a group conversation, turning towards the active speaker, it leads to more natural-looking movements than when the trajectories and motions are generated from a baseline. Imitated movement thus improves human-AI interaction. Imitation of affective expressions by a virtual agent displaying a positive expression right after a smile has been detected, brings about a positive emotion in the user, and effectively leads to anthropomorphism (Numata, et al., 2020). Adaptation is also possible at the level of personality. PhD researcher Laduona Dai, supervised by Merel Jung, Max Louwerse and me, currently examines if aligning the personality profile of a chatGPT-based VR pedagogical agent to the profile of the student results in better interactions and, ultimately, improved learning outcomes.

A more complex case study is that of a human-AI team currently developed in the STEADFAST consortium for which I serve as primary investigator with my colleague Renato Calzone as project manager. In this project, we aim to create human-swarm teams that consist of trained operators in charge of a group of minuscule drones with different levels of autonomy. The goal of the team is to support situational awareness in adversarial conditions. These can include civilian scenarios, such as natural disasters or industrial accidents, as well as military scenarios. The drones are able to operate fully autonomously in a coordinated manner under conditions where the connection to the human operator might be lost or where the operator is unable to contribute in an optimal way due to cognitive overload. Unlike in the case of FPV-drones where the operator is in control unless specific circumstances require otherwise, drone

swarms, particularly in environments with GNSS-denied, need to be able to execute operations without human guidance, authorization or notification. Making use of advances in neuromorphic computing, the swarming drones should have the ability to self-navigate and self-position, so that they can collect information efficiently and safely. Having learned in simulated environments, autonomous swarms can decide how to distribute their resources and control their parameters (e.g., airspeed or altitude) to meet their objective.

This set-up is referred to as human-out-of-the-loop. It is a more extreme scenario than operations where the operator is integrated in the system's control loop (human-in-the-loop) or where they monitor the operations of the swarm and can intervene when necessary (human-on-the-loop). Adapting levels of autonomy is challenging to the operator and requires some form of shared awareness (Hauptmann et al., 2023). In the STEADFAST project, which is a collaboration of 30 academic, industrial and governmental partners, we develop different simulated scenarios regarding the team mission in which the cognitive complexity of the assignment is manipulated. With the help of sensors measuring EEG, heart rate, respiration, electrodermal activity and eye movements, we collect information about the cognitive state of the operator in real time. Drowsiness and mental fatigue, for example, caused by prolonged high workload are reflected in patterns of pupil dilation and the brain frequency bands. This data is subsequently used in models of flexible and adaptive swarm autonomy to determine optimal levels of human control in relation to task complexity.

As you can imagine, this goal raises a lot of questions regarding legal and ethical responsibility, trust and organizational acceptance. This research is very important in all conditions where the use of AI agents is needed to collect and process large amounts of data quickly. Next to the STEADFAST project which focuses on situational awareness, similar issues arise in the context of cyber incident response and other areas related to hybrid warfare. These are all topics on which we collaborate with Roy Lindelauf and his team from the Data Science Center of Excellence of the Netherlands Defence Academy, with partners in the ELSA Lab Defense, as well as with Maureen Sie and other philosophers of ethics at Tilburg School of Humanities and Digital Sciences and at Tilburg Institute for Law, Technology and Society.

5. Cognitive Warfare

When everyone believes you
What's that like?

- Taylor Swift, The Man (2019)

Spreading mis- and disinformation involves collaboration between AI and human users who intentionally or unknowingly participate in the process through engaging with and sharing false and manipulated content. This is sometimes described as 'participatory disinformation'. In 2024, the World Economic Forum in their report listed mis- and disinformation as the most severe risk in the next two years. In January 2024, the European Union published its second report on foreign information manipulation and interference threats (FIMI) showing their impact.[5] The organizations most often targeted by FIMI attacks were the European Union, NATO, the armed forces of Ukraine, the UN, and various media organizations. Gender-based and anti-LGBTIQ+ FIMI attacks revealed a worrying trend, with a rise in hate crimes against members of the community across Europe as a likely consequence.

FIMI actors typically attempt to hijack the attention created by certain events to shift the narrative related to the event. What is the role of AI in this nefarious effort? We know since Cambridge Analytica how effective microtargeting can be achieved by training content recommendation algorithms on massive amounts of online behavioral data. GenAI can be used to create colossal numbers of fake online websites spreading false content through search engines. It can also be employed to quickly create content and narratives that are in line with the intended goal of the disinformation campaign. For example, in Taiwan's presidential elections this year, a 300-page ebook most likely written by genAI was created to spread false allegations that Taiwan's president came to power through sexual promiscuity. The message was subsequently amplified by avatars appearing on TikTok, YouTube and Instagram as newscasters reading passages from the book. This is not an isolated example. In 2023, a synthetically altered video encouraged Ukrainians to participate in a coup. Yet another example of AI-generated content is the use of a deepfake audio message during the 48-

5 https://www.eeas.europa.eu/sites/default/files/documents/2024/EEAS-2nd-Report%20on%20FIMI%20Threats-January-2024_0.pdf

hour moratorium ahead of the polls in Slovakia this year. In the audio, which showed clear signs of AI manipulation, the liberal Progressive Slovakia party candidate Michal Šimečka appeared to be discussing with a journalist from the independent newspaper Denník N how to rig the elections. In the UK, an audio clip emerged of the current prime minister Sir Keir Starmer supposedly swearing at a member of his staff. One could not agree more with Gordon Crovitz from NewsGuard, a company that analyses online content for trustworthiness, that "[chatGPT] is [...] the most powerful tool for spreading misinformation that has ever been on the internet" (Hsu & Thompson, 2023). Even if these and similar efforts did not achieve any serious direct effect so far, in an information space saturated with deepfakes, citizens start doubting any content shared by public media, resulting in a 'liar's dividend'.

We have seen a growing effort within the AI community to design tools that serve to detect disinformation activities on online platforms and search engines. Here in Tilburg, PhD researcher Timo Westlake, under the supervision of Roy Lindelauf, Boris Čule and myself, works on the development of an early detection algorithm for different types of disinformation on Twitter-X by using features pertaining to information propagation. Thanks to funding we obtained from the European Media Information Fund, my colleagues from Tilburg Algorithm Observatory Henry Brighton, Chris Emmery, Ronja Rönnback and me collaborate with the Data Science Center of Excellence of the Dutch Defense, Charles University and Filip Milde from Prague Pride to monitor and analyze the spread of anti-LGBTQ propaganda in a network of web domains across Europe. By leveraging these and other advanced AI techniques, we can develop robust systems to mitigate the spread of disinformation online. These efforts can significantly strengthen the overall integrity of online discourse.

6. Closing remarks

As a quote for this inaugural lecture, I chose a fragment from Gustav Meyrink's novel The Golem. According to a Jewish legend, the golem – an artificial living being - was created by rabbi Löew in Prague in 1580. The golem was manufactured from clay at the banks of the Moldau River and brought to life to save Jewish people from danger. The golem had no consciousness and thus no moral obligations; if he committed a crime, its creator was responsible. I am sure we can think of many parallels between this legend and the use of artificial intelligence today. However, in Meyrink's modernist fantasy, the focus is not on the use of an artificial being for physical protection. Rather, it is on the fuzzy distinction between the "real" world as opposed to dreams and hallucinations, the uncertainty of individual and collective identity and the fragmentation of the subjective conscious experience. "[Our] soul is [...] composed of many 'selves', just as a colony of ants is composed of many single ants," Meyrink writes. The main protagonist of his novel faces the challenge of discovering and integrating these selves, his journey symbolized by the tarot card of The Fool – the seeker and the trickster who accompanies the beginning of each new spiritual cycle captured by the major arcana. The quote from The Golem refers to the impossible effort that we are engaged in, trying to understand the unifying essence of human consciousness while being unable to step outside of it. I would like to believe that the creation of artificial intelligence can offer us the tools to transgress the boundaries of our individual and collective cognitive system and thereby finally understand its workings.

7. Acknowledgements

Oh, I'm a lucky man
To count on both hands
The ones I love

- Pearl Jam, Just Breathe (2009)

We have now come to the final part of this lecture and possibly the most important one, namely the part where I would like to thank everyone who was instrumental in my journey to where I stand today. If this journey would be described in philosophical terms, it would lead from Wittgenstein to Foucault, to Schopenhauer and bits of Kafka's Castle thrown in for fun. In the spirit of Joscha Bach, I often remind myself that I am not a person, I am a piece of software running on the brain of a random ape for a few decades. For what it's worth, hereby my sincere thanks as a random ape.

Let me first thank the Executive Board of Tilburg University, the Rector and the Dean of the Tilburg School of Humanities and Digital Sciences for appointing me to this position.

To the wonderful teachers and researchers who inspired me throughout my academic journey so far. The philosophers, Old English scholars and poets at Charles University. To my study friends "The Coffin Makers". The linguists at the University of Washington and University of Tromsø. The ILLC people: Marc, who made me love and hate modal logic (particularly when he refused to do the proofs for me), Robert and Balder, for writing the first scientific papers with me, and to all the senior staff for creating a unique environment of intellectual freedom and equality for the PhDs at the institute.

At the Department of Cognitive Science & AI at Tilburg University: Eric, Pieter, Max, Afra, Grzegorz, Emmanuel, and Drew - I have worked together with you for more than 10 years now which is a lot considering that during the same period we built three new educational programs, the biggest department at this university, dealt with incredible staff shortages, and recently, periods of research investments followed by dramatic cuts and calls to curb internationalization at Dutch universities. Every time

I think we are out of the woods, there is another challenge waiting just around the corner. To my trusted colleagues Eva and Karin, the heart of our department. My fellow heads of department at TSHD. To Loes, the best example of female leadership I know, and to Robin, who is the HR advisor we all need in our lives. All my wonderful past and present colleagues: Marjolijn, Dries, Jaap, Joke, Jolien, Ad, Aske, Henry, Travis, Çiçek, Marijn, Renato, Roy, Harm, Noortje, Myrthe, Koen, Görkem, Murat, Giacomo, all the wonderful PhD students I have been lucky to work with... the list is long, but you know who you are!

Finally, I want to thank my friends and family.

To those who are no longer with us but will never be forgotten.

To my daughter's grandparents, Albert and Agnes, for their kind support throughout the years.

To my best friends, Zuzana and Alexander. We don't see each other often but when we do, I know there is no one better out there.

To my mother, for her special brand of common sense and pragmatism.

To my sister, for her contagious energy and perseverance.

To Wouter, Aimée and Tommie, for being the most amazing and wonderful kids and adults, each one of them in their own way, for taking care of one another and for dealing with what life serves them with an admirable levelheadedness and positivity.

Most importantly, I want to thank Eric, for his love, acceptance, wisdom, and unconditional kindness. He believes in me even when no one else does.

I have spoken

8. References

Ardiel, E.L., & Rankin, C.H. (2010). An elegant mind: learning and memory in Caenorhabditis elegans. *Learning & Memory, 17*, 191-201.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., et al. (2018). The Moral Machine Experiment. *Nature 563(7729)*, 59-64.

Bayne, T., Seth, A.K., Massimini, M., Shepherd, J., Cleeremans, A., Fleming, S.M., ...& Mudrik, L. (2024). Tests for consciousness in humans and beyond. Trends in Cognitive Sciences.

Breazeal, C., & Picard, R. (2006). The role of emotion-inspired abilities in relational robots. *Neuroergonomics: The Brain at Work*, 275-292.

Broers, A., Louwerse, M.M., & Postma, M. (2020). Creating ambassadors of planet Earth: The Overview Effect in K12 education. *Frontiers in Psychology, 11*, 540996.

Burda, R. (2023). *Cognitive warfare as part of society: Never-ending battle for Minds.* Hague Centre for Strategic Studies.

Chambers, D.W. (1983). Stereotypic images of the scientist: The draw-a-scientist test. *Science Education, 67(2)*, 255-265.

Damasio, A. (1994). *Descartes' Error: Emotion, Reason and the Human Brain*. Pan Books, London.

Damasio, A., & Damasio, H. (2024). Homeostatic feelings and the emergence of consciousness. *Journal of Cognitive Neuroscience, 36(8)*, 1653-1659.

Dias Da Silva-van Riel, Gonçalves, Ó.F., Branco, D., & Postma, M. (2022). Revisiting consciousness: Distinguishing between states of conscious focused attention and mind wandering with EEG. *Consciousness and Cognition, 101(18)*, 103332.

D'Mello, S.K., Tay, L., & Southwell, R. (2022). Psychological measurement in the information age: Machine-learned computational models. *Current Directions in Psychological Science, 31(1)*, 76-87.

Garfinkel, S.N., Schulz, A., & Tsakiris, M. (2022). Addressing the need for new interoceptive methods. *Biological Psychology, 170,* 108322.

Glowacki, D.R., Williams, R.R., Wonnacott, M.D., Maynard, O.M., Freire, R., Pike, J.E., & Chatziapostolou, M. (2022). Group VR experiences can produce ego attenuation and connectedness comparable to psychedelics. *Scientific Reports, 12(1),* 8995.

Graziosi, M., & Yaden, D. (2021). Interpersonal awe: Exploring the social domain of awe elicitors. *The Journal of Positive Psychology, 16(2),* 263-271.

Hauptman, A.I., Schelble, B.G., McNeese, N.J., & Chalil Madathil, K. (2023). Adapt and overcome: Perceptions of adaptive autonomous agents for human-AI teaming. C*omputers in Human Behavior, 136,* 107451.

Hawkins, R.X., Goodman, N.D., & Goldstone, R.L. (2019). The emergence of social norms and conventions. *Trends in Cognitive Sciences, 23(2),* 158-169.

Hollander, G. de, Forstmann, B.U., & Brown, S.D. (2016). Different ways of linking behavioral and neural data via computational cognitive models. *Biological Psychiatry, 1(2),* 101-109.

Lee, D. (2013). Decision making: From neuroscience to psychiatry. *Neuron, 78(2),* 233-248.

Mattheij, R.J.H., Postma, M., & Postma, E.O. (2015). Mirror, Mirror in the Wall: Is there mimicry in you all? *Journal of Ambient Intelligence and Smart Environments, 7(2),* 121-132.

Melloni, L., Mudrik, L., Pitts, M., & Koch, C. (2021). Making the hard problem of consciousness easier. *Science, 372(6545),* 911-912.

Michel, F., & Gandon, F. (2024). Pay attention: a call to regulate the attention market and prevent algorithmic emotional governance. *arXiv:2402.16670.*

Nagel, T. (1974). What is it like to be a bat. *The Philosophical Review,* 435-450.

Picard, R. (1997). *Affective Computing*. The MIT Press, Cambridge, Massachusetts.

Postma-Nilsenová, M., Brunninkhuis, N., & Postma, E.O. (2013). Eye gaze affects vocal intonation mimicry. In: Knauff, M., Pauen, M., Sebanz, N., & Wachsmuth, I. (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society, pp. 1139-1144.

Postma-Nilsenová, M., & Postma, E.O. (2013). Auditory perception bias in speech imitation. *Frontiers in Psychology, 4*, 826.

Read Montague, P., Dolan, R.J., Friston, K.J., & Dayan, P. (2012). Computational Psychiatry. *Trends in Cognitive Sciences, 16(1)*, 72-80.

Rombout, L., & Postma-Nilsenová, M. (2022). Disentangling the effects of matching content and simultaneous speech on phonetic adaptation. *Auditory Perception & Cognition, 5(1-2)*, 107-128.

Schneider, P., Sluming, V., Roberts, N., Scherg, M., Goebel, R., Specht, H.J.,..., & Rupp, A. (2005). Structural and functional asymmetry of lateral Heschl's gyrus reflects pitch perception preference. *Nature neuroscience, 8(9)*, 1241-1247.

Seth, A.K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience, 23(7)*, 439-452.

Shteynberg, G., Hirsh, J.B., Garthoff, J., & Bentley, R.A. (2022). Agency and identity in the collective self. *Personality and Social Psychology Review, 26(1)*, 35-56.

Shteynberg, G., Hirsh, J.B., Wolf, W., Bargh, J.A., Boothby, E.B., Colman, A.M.,...& Rossigna-Milon, M. (2023). Theory of collective mind. *Trends in Cognitive Sciences, 27(11)*, P1019-1031.

Smallwood, J., & Schooler, J.W. (2015). The science of mind wandering: Empirically navigating the stream of consciousness. *Annual Review of Psychology, 66(1)*, 487-518.

Tsoumani, O., & Postma-Nilsenová, M. (2013). Perceiving sounds: analytic and synthetic listening, global-local processing and possible links with empathy and self-construal. In: Knauff, M., Pauen, M., Sebanz, N., & Wachsmuth, I. (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society, pp. 3587-3592.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science 185(4157),* 1124-1131.

van Limpt-Broers, H.A.T., Postma, M., & Louwerse, M.M. (2024). Measuring transformative virtual reality experiences in children's drawings. *Memory & Cognition*.

van Limpt-Broers, H.A.T., Postma, M., van Weelden, E., Pratesi, S., & M.M. Louwerse (2024). Neurophysiological evidence for the overview effect: a virtual reality journey into space. *Virtual Reality*.

Yaden, D.B., Haidt, J., Hood Jr, R.W., Vago, D.R., & Newberg, A.B. (2017). The varieties of self-transcendent experience. *Review of General Psychology, 21(2)*, 143-160.

Colophon

design
Beelenkamp ontwerpers, Tilburg
photography cover
Maurice van den Bosch
layout and printing
Studio | powered by Canon