Detecting misinformation and disinformation with AI

Dr. Petra Aczél, Moholy-Nagy University of Art and Design, Budapest

Is there an everyday life, let alone a future, without artificial intelligence (AI)? And is there any kind of communication without lies? – These were the questions the module discussion started with. While the answer to the first question was an almost yes, because there always will be laggards and drop-outs, as present digital divides prove, the answer to the second is clearly no. Falsifying and lies seem to be necessary parts of social communication. Recognizing this, however, does not mean that we are correct in our own human ability to identify a lie. We tend to believe that we are almost 100 percent successful in detecting when someone is misrepresenting information, when even the best debunkers are less than 60 percent successful. Why? Because we are more deceptive in interactive situations. When access to information doesn't require any special effort, we are more deceptive. When we feel the content of the message close to us, when we resonate with it, we are more prone to be derailed from the truth. How much better is AI than us in this respect? On the one hand, the current input of AI "knowledge" is still partly human content. So the starting point can be false. On the other hand, AI's capacity allows it to source-check, identify patterns of error, realize and predict the path of information spread and, based on that, provide faster, more accurate answers.

In this module, we have used a social scientific and explorative approach to answer the ways in which information can be faked, what differences may be identified between mis- and disinformation. We talked about the AI programs that can be created and accessed based on this, and the ways in which human attention needs to evolve in parallel.

As a starting point, we analyzed our view of AI. We started from the so-called Rumsfeld-matrix, according to which there are some things in the world that we know that we know. For example, that humans need air to survive. Then there are the phenomena that we know we don't yet know. An example of this is, say, our picture of deep-sea life, because we recognize that we don't yet fully know it. There are also things that we do not know that we actually know. We do not identify our intuition as knowledge, for example, and tend to dismiss it and rely on information alone. But we also need the everyday wisdom that we have not acquired through learning, what the Hungarian economist-philosopher Mihály Polányi called 'tacit' knowledge. But sometimes we don't even know we have it. Finally, events and phenomena can also be classified in the category of things we do not know we do not know. This is a statement that was highlighted and much debated and discussed in Donald Rumsfeld's response to a press conference in 2002. When we do not know that we do not know something, it does not mean that we do not talk about it. Often the opposite is true. We talk about it a lot, precisely because we do not recognize our ignorance.

In a sense, we do the same with AI. And talking about also has an impact on the world of business and policy making, and is becoming a business in itself. We can identify at least four styles in the AI discourse. The first is the 'utopian'. This narrative paints a complex vision of the future, in which AI appears as a central drive that will change our everyday and professional lives. In these discourses, humans and technology coexist, the former has accepted that there is something that has the capacity to act more efficiently than humans can do, and is searching the place of humanity in relation to it. This style has a major impact on business and IT competition. The second style is different in tone, much darker, indeed. You could call it 'apocalyptic'. It implies that AI is the evil alien that outgrows and, in a way, robs us of our independence, that occupies our place, influences

our social relations and autonomy, and that gains power over us. It is the style that appeals most to policy-makers and regulators, legislators. The third style can be labelled as ‘mystical’ or ‘enigmatic’, and it bears the characteristics of scientific discourses. It is suggestive rather than assertive, it relies on data, it is cautious, it allows for multiple inferences. Scientific discussions draws on this style. Finally, the ‘pragmatic’ style speaks the language of technologists. It is a descriptive way of talking about what AI can do in its current state, what can be improved or should be improved. This style can easily fuel innovation in IT. While the first assumes that we know that we know what AI is, the second proposes that we don't know that we know. The third suggests that we know that we don't know, and the fourth displays the belief that we know that we know. However, none of them apply the principle of the ‘we don't know that we don't know’. This is important because we are in a similar situation with falsified information. At times we are overconfident that they are detectable, at other times too anxious about their effects, at still other times we tend to be uncertain about their implications.

In the course of the discussion, we discussed how the latest polls show that fake news is considered to be one of the top three risk for the next two years, considered by many as a more serious problem than war. For this reason, we looked at the main methods of falsification, such as concealment and its techniques (silencing, decontextualization, omission, erasure, non-referencing) and falsification and its procedures (exchange of facts, wrong resource and reference, pseudo/false experts, biased language). In light of this and the spread of fake news, we discussed four different fake information identifier AI programs that can also act as web browser extensions. Each of them uses a different logic to search for and detect fake information. One examines the text in a large database, one considers the spreading patterns, one looks at the contexts, and one starts off from the production of the fake news, suggesting that imitation facilitates detection as well. Composing genre of fake information deepfakes were also discussed in the context of increasingly powerful visual forgery techniques and their potential uses.

Finally, we talked about the principles of the future: the skills and attitudes that can support and enable people to remain self-aware in detecting fake news, and continue to program the principles of detection wisely, promote AI-AI learning within this frame, while staying humble in realizing what we can and what we cannot know.