Cognitive biases in humans and machines: Interview with Helena Matute

Luis Cásedas
Dept. de Psicología Básica, Universidad Autónoma de Madrid, España

Tipo de artículo: Entrevistas, Multilingüe.
Disciplinas: Psicología, Inteligencia Artificial.
Etiquetas: sesgos cognitivos, toma de decisiones, razonamiento, ayudantes inteligentes.

Helena Matute is Professor of Experimental Psychology at the University of Deusto. Recently appointed Full Member of the Spanish Academy of Psychology, her research has significantly advanced our understanding of cognitive processes such as learning and memory, particularly with regard to their biases. In this interview, I talk to Dr. Matute about cognitive biases: their origin, nature and implications. We also discuss the presence of biases in artificial intelligence, and how it could perpetuate and amplify human errors. The interview concludes with Dr. Matute's reflections on the need for regulatory frameworks to ensure the ethical and safe development of this technology.

Question – What are cognitive biases?

Answer – They are systematic and predictable errors that happen to most people in a similar way, and that occur in cognitive processes such as attention, learning, memory, reasoning, or decision-making.

Q – Why do we have cognitive biases (rather than an infallible mind)?

A – As a result of natural selection, we are designed to adapt to the world, not to be perfect. If there are tricks and shortcuts to help us make quick and effective decisions, we will use them. These shortcuts are what we call heuristics, and biases are their negative byproduct. Heuristics allow us to make quick decisions, which are often correct, even when we have little information, which makes things easier for us. It would often take a lot of time and energy, which we do not always have, to act rationally from all angles. This is why we use heuristics, which generally work very well. The problem is that these shortcuts are not very rational but have been selected to be effective in specific situations, and so are very likely


(cc) Helena Matute.

CC BY

to lead us to the wrong answer in situations other than the usual ones. Broadly speaking, this would be the reason why biases arise.

Q – Could you describe some of our most common cognitive biases?

A – There are so many biases. We could talk at length, even if we were to focus only on the most common ones, such as confirmation bias, availability bias, consensus bias, causality bias... The latter, causality bias, is the one we have worked on most in our research team (e.g., Matute et al. 2015; Matute et al. 2019). The causality bias is the mistaken belief that one event is the cause of another, which often occurs when the possible cause and the possible effect happen consecutively in time. Imagine you have a headache, and you take an alternative medicine that has been recommended to you, but that has no real effect. If you feel better the next day, it would not be unusual for you to attribute the improvement to the medicine, even though it may have been caused by something else. You have developed a cause-effect bias, which can sometimes have serious consequences, for example, when someone stops taking a treatment that could actually cure them because they trust an ineffective medicine.

Q – Is there anything we can do to minimise our cognitive biases?

A – The most important thing is to be aware of them. If we are aware that we have them, and in important situations make the effort to stop and think and act slowly, I think they can be minimised considerably. For example, those of us who are involved in research do so, at least, when we are at work, by applying what is known as the scientific method. Inspired by this idea, our team has recently developed a project in schools across Spain, in collaboration with the Spanish Foundation for Science and Technology (Martínez et al., 2024). The basic idea is to teach students, from a very young age, to be aware of some of their own biases and to internalise the scientific method as a tool for critical thinking. Best of all, it works! We found that the students who took part in this workshop showed greater resilience against causality bias than their peers who did not. We also found that this effect is not only visible at the end of the workshop, but is maintained for six months after the workshop. These results encourage us to think that it is also possible to work with other biases in the educational system to help young people acquire the necessary tools to minimise their impact.

Q – Let's now talk about artificial intelligence (AI) and its biases. In general terms, how do AI systems work? Can you give some examples?

A – In essence, they are machines that learn from data. This learning process can take place in a number of ways. One of these is reinforcement learning, where the AI adapts its behaviour as it interacts with users. For example, every time we click on a video recommended by a social network, we are reinforcing the AI algorithm, which learns what kind of content to show us in the future to capture our attention and keep us on the platform longer. There is also supervised learning, where AI is trained through specific cases (e.g., by teaching it the correct answers to specific questions), and unsupervised learning, where AI identifies patterns present in large amounts of data (e.g., by clustering similar information without explicit instructions on how to do so). Prominent among these systems are Large Language Models, such as ChatGPT, which have become very popular in the last year due to their ability to generate apparently coherent text.

Q – Do AIs inherit our biases as they learn?

A – That's right, AIs acquire our biases. This can be through the people who design them, but also through the databases that are used to train them, as well as through their interaction with the people who use them.

Q – What kinds of biases are common in AIs?

A – They often develop discriminatory biases, such as race or gender bias, which in turn are based on other more general cognitive biases, such as causality or representativeness bias. For example, an AI trained with a hospital database that is biased (e.g., because white people have traditionally received more and better treatment in that hospital than people of other ethnic groups) could end up concluding that white people have a greater need or urgency for treatment. The problem, moreover, is that these biases are always detected after the fact. We can't know if the AI used by our insurance company, for example, is biased against women until

CC BY

one day a news story comes out saying that someone suspected that particular bias in that particular AI, reported it, and it turned out to be true. We need to be aware that this will continue to be the case, and be forewarned. We cannot assume that AIs are neutral and objective, because they are not.

Q – Can humans, in turn, inherit biases from AIs?

A – Yes indeed, and this is something we are also testing in our lab. Currently, the law says that when AI is used in high-risk decision-making contexts (e.g., medical diagnosis or legal decisions), there must always be a responsible person in the process to ensure that the decision is correct and free of bias. The problem, however, and what we see in our studies, is that people who work with biased AIs end up being very susceptible to their biases and may end up reproducing them. In fact, our results show that after working briefly with a biased AI, people continue to reproduce the same errors even when the AI is no longer present (Vicente & Matute, 2024).

Q – The risks of AI could go beyond those related to its biases, which some experts, including you, also warn about. However, some argue that this fear stems precisely from a cognitive bias, in this case a human one, whereby we overestimate the danger of the emergence of previously unknown technologies. We have seen this in the past with various technological advances (e.g. cars, X-rays, electricity). Why is AI different?

A – I think a key aspect is regulation. When cars were invented, we started to develop regulations to control their use so that we could take advantage of their positive aspects while minimising their risks. Today, cars are under control, and yet we are still adding rules to regulate their use. Similarly, the use of X-rays in hospitals is subject to high safety protocols, as is the electrical installation in any house. The same should apply to AI. In this case, however, we should go further, because another key aspect has to do with the scale of its impact. I compare AI to atomic energy: a very powerful technology that, if used well, can be a huge advance for humanity; but if used badly, it has the potential to destroy it. Moreover, it is a technology that is advancing at a breakneck pace, which provide an additional incentive to regulate it without delay. If we leave the development of this powerful technology in the hands of companies interested in exploiting it for their own benefit, as has largely been the case to date, humanity may soon find itself in very serious trouble. That is why we need to regulate it, and we need to do it now.

References

Martínez, N., Matute, H., Blanco, F., & Barberia, I. (2024). A large-scale study and six-month follow-up of an intervention to reduce causal illusions in high school students. *Royal Society Open Science*, 11, 240846.

Matute, H., Blanco, F., & Díaz-Lago, M. (2019). Learning mechanisms underlying accurate and biased contingency judgments. *Journal of Experimental Psychology: Animal Learning and Cognition*, 45, 373-389.

Matute, H., Blanco, F., Yarritu, I., Díaz-Lago, M., Vadillo, M. A., & Barberia, I. (2015). Illusions of causality: How they bias our everyday thinking and how they could be reduced. *Frontiers in Psychology*, 6:888.

Vicente, L., & Matute, H. (2023). Humans inherit artificial intelligence biases. *Scientific Reports*, 13, 15737.

Further reading

Matute, H. (2019). *Nuestra mente nos engaña: Sesgos y errores cognitivos que todos cometemos*. Shackleton books.

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., … Mindermann, S. (2024). Managing extreme AI risks amid rapid progress. *Science*, 384, 842-845.

Contact the authors

Luis Cásedas: luis.casedas@gmail.com; Twitter/X: @lcasedas

Helena Matute: matute@deusto.es; Twitter/X: @HelenaMatute

Manuscript received on October 3rd, 2024.
Accepted on November 3rd, 2024.

This is the English version of
Cásedas, L. (2024). Sesgos cognitivos en humanos y máquinas: Entrevista con Helena Matute. *Ciencia Cognitiva*, 18:3, 43-46.

CC BY