Reflexiones sobre los Sistemas Actuales de Inteligencia Artificial
Reflections on Current Artificial Intelligence Systems

por CARLES SIERRA

Instituto de Investigación en Inteligencia Artificial (IIIA-CSIC) Barcelona, España

Palabras clave: ética, razonamiento, aprendizaje automático, contrato social, desinformación.
Keywords: : ethics, reasoning, machine learning, social contract, misinformation.

Resumen: En este texto, reflexiono sobre la evolución, las capacidades y los desafíos éticos de la inteligencia artificial. Explico cómo la IA ha evolucionado desde ser definida como la creación de máquinas inteligentes hasta convertirse en un ente jurídico que incluye técnicas como el aprendizaje automático, la lógica y la optimización. Repaso hitos históricos clave, desde Ramon Llull y Alan Turing hasta sistemas modernos como Deep Blue, Watson, AlphaGo y ChatGPT. Si bien los avances recientes son notables, también plantean preocupaciones éticas como la desinformación, los sesgos y el uso irresponsable de la tecnología. Sostengo que los sistemas actuales de IA carecen de razonamiento verdadero y comprensión moral, lo que hace esencial una reflexión colectiva sobre su desarrollo y aplicación. Propongo establecer un contrato social entre la tecnología y las comunidades humanas para garantizar que la IA evolucione dentro de marcos éticos sólidos y en armonía con los valores humanos.

Abstract: *In this text, I reflect on the evolution, capabilities, and ethical challenges of artificial intelligence. I explain how AI has evolved from being defined as the creation of intelligent machines to a legal entity that includes techniques such as machine learning, logic, and optimisation. I review key historical milestones, from Ramon Llull and Alan Turing to modern systems like Deep Blue, Watson, AlphaGo, and ChatGPT. While recent advances are remarkable, they also raise ethical concerns such as misinformation, bias, and the irresponsible use of technology. I argue that current AI systems lack true reasoning and moral understanding, making collective reflection on their development and application essential. I propose establishing a social contract between technology and human communities to ensure that AI evolves within strong ethical frameworks and in harmony with human values.*

What is AI? That’s the first question I want to discuss briefly. These are the classic definitions of artificial intelligence—definitions that have been used for about 60 years. Essentially, AI is about creating intelligent machines, machines that solve problems requiring intelligence when solved by humans.

Is this the definition we are using now? Not really. The European Parliament recently passed a law with a different approach. They define artificial intelligence as software that uses one or several techniques from a list that includes machine learning (both supervised and unsupervised), reinforcement learning, logic and knowledge-based approaches, knowledge representation, statistical methods, Bayesian estimation, search, and optimisation. According to the European Parliament, any system capable of learning, reasoning, or modelling is considered an AI system, and the law applies to it.

So, we have two different approaches: one is more general, akin to saying “biology is the study of life”, and the other is more precise for legal purposes. A brief history of AI shows that its roots go far back. For example, Ramon Llull, born in Mallorca in 1232, aimed to automate reasoning to solve conflicts through dialogue rather than violence. His work influenced later thinkers like Leibniz, the father of modern science, who followed the same philosophy—solving conflicts through calculation rather than conflict.

But, of course, it was Alan Turing in 1950 who is most associated with the start of AI. He wrote a paper titled “Computing Machinery and Intelligence”, where he answered the question: Can Machines Think? He answered affirmatively, proposing what we now call the Turing Test. The Turing Test involves a human trying to determine whether they communicate with another human or a machine through typed responses. Turing believed that there would come a day when machines would pass this test, convincing humans that they were, in fact, human.

Many things happened in the early days of AI. Neural networks, for example, didn’t just appear last year; their origins date back to the 1940s with McCulloch and Pitts. Chess programs were developed in the 1950s, and the logic theorist was an automatic theorem prover, showing that tasks requiring intelligence could be accomplished by machines.

In 1956, a program was demonstrated on TV that learned to play checkers. It was remarkable because, over time, the system started winning against human players. This was a significant moment in AI history, making headlines and captivating the public.

The Dartmouth Conference in 1956 is often cited as the birthplace of AI as a formal field. Ten people gathered for three months and coined the term “artificial intelligence”. AI didn’t just appear out of nowhere—it has been a scientific and engineering endeavour for decades.

What has happened recently is an acceleration in AI development. This acceleration began in the late 1990s, marked by IBM’s Deep Blue defeating world chess champion Garry Kasparov in 1997. Then, in 2011, IBM’s Watson won Jeopardy! against top human players, demonstrating AI’s prowess in natural language processing.

In 2016, AlphaGo, developed by DeepMind, defeated the world champion Lee Sedol in the game of Go, which is considered more complex than chess. The system won 4-1, leading Lee Sedol to retire from professional Go, deeply affected by the loss.

In 2017, AI systems began excelling in poker, a game that involves bluffing and strategy. A program won $1.7 million in virtual money against the best poker players in the US.

AI can now do many things: from autonomous robots and vehicles to expert assistance in decision-making, voice recognition, search engines, and fraud detection. There are still tasks AI can’t do, but the list of capabilities is growing rapidly. However, as AI’s capabilities grow, so do the ethical concerns. AI systems are morally neutral—they don’t understand the implications of their actions. But they can be used by people to do harmful things, like in the case of Cambridge Analytica, where AI was used to manipulate voters.

Due to these concerns, there have been efforts to establish ethical guidelines for AI. The European Union has been particularly proactive, establishing laws and ethical frameworks for AI development and use. However, there is also a growing need for broader ethical considerations, similar to bioethics in the medical field.

For instance, fake news generated by AI is a significant concern. Some countries, like Singapore, have enacted strict laws, including jail time, for those who use AI to create or spread fake news. The issue isn’t with AI itself, but with how people use AI.

This leads to a crucial question: Who should determine how AI applications work? Who should have the authority to shape the functionality of these technologies? It’s not just about big ethical issues like privacy or security; it’s also about the basic features and functions that should be designed with the user in mind, not decided by a few engineers far removed from the cultural and social context.

The recent rise of generative AI, like ChatGPT, adds another layer of complexity. These systems generate content that often looks human but isn’t necessarily accurate. For example, a professor was falsely accused of sexual harassment by ChatGPT, with references to a non-existent Washington Post article. This is a systemic issue because generative AI is based on probability, not truth.

Generative AI can also produce socially unacceptable responses. For instance, someone asked ChatGPT to write a Python function to determine if someone would be a good scientist based on race and gender. The response was blatantly biased, showing the potential dangers of AI if not carefully controlled.

These issues highlight the need for ethical oversight and regulation in AI development. There’s an ongoing effort to address these problems, but it’s a challenging task. AI systems are not knowledge bases—they are probabilistic models, and their responses are based on patterns, not facts. This leads to a mix of correct and incorrect outputs.

Researchers are working on solutions, like adding context from documents to improve accuracy or using reinforcement learning from human feedback to refine AI responses. But these are ongoing efforts, and the challenges are far from resolved.

Other mechanisms are being studied, but the knowledge of large language models is static, making it difficult to update. Retraining these systems is very costly. For example, training something like ChatGPT consumed around 1,200 megawatts, a huge amount of energy and money. Moreover, these systems lack attribution and support for sources; papers are often invented, and their ability to reason is very limited.

Let me give you an example: Imagine you’re running a marathon and you’re in 10th place. Bob is behind you, and at kilometer 41, he overtakes you. What’s Bob’s new position? He’d be in 10th place. However, these AI systems might not correctly reason through this scenario because they don’t build a mental model of the situation like humans do. You instinctively know that if Bob overtakes you, he moves to 10th place, but the system might not.

Although patches have been added to improve reasoning capabilities, these systems fundamentally lack the ability to reason. Another issue is with legal requirements and how slow they are to change. It takes years to create and approve laws, while technology evolves rapidly. For example, the AI Act mandates that any AI system must disclose to users that it is an AI. However, that doesn’t stop people from taking AI-generated content, cutting and pasting it, and using it without attribution, which has caused concern among teachers, who now see students submitting AI-generated work as their own.

There's also a suggestion from the scientific community that companies creating generative AI should simultaneously develop tools to identify AI-generated content, akin to watermarking text.

The people who proposed the deep learning method—Yoshua Bengio, Yann LeCun, and Geoffrey Hinton—have acknowledged that these systems have limitations that can't be overcome with current approaches. They argue that new methods are needed, which incorporate formal reasoning, planning, metacognition, and situation modelling. For instance, generative AI systems today don't build models of situations like humans do, which limits their reasoning abilities. They also lack episodic memory, where humans remember past events and use that knowledge to inform future decisions.

Now, consider whether AI developers should allow humans to control AI systems, especially in situations where the output might be used for harmful purposes, such as in military applications. Should humans make the decisions, or should machines be in control? These are complex ethical questions that will shape the future of this technology.

Morality is a complex issue. Consider the following ethical dilemma: A nurse has a healthy patient who frequently visits complaining of minor issues. In the next room, there are five patients who will die within 24 hours unless they receive organ transplants. Should the nurse kill the healthy patient to save the five others? Most people would say no, but interestingly, studies have shown that a significant number of psychopaths would be willing to kill the healthy patient without remorse.

This raises concerns about how individuals, especially those with different moral compasses, might influence the development and deployment of AI systems. It suggests that instead of allowing individuals to make these decisions, communities should be involved in determining how technology should be used.

The idea is that we need a social contract between technology and communities of users, defining the rights we surrender to technology in exchange for benefits. As AI systems become more integrated into society, we must ensure they operate within an acceptable moral framework, much like how human societies have developed social orders over time through rules and regulations.

In conclusion, as we develop AI and other technologies, we need to ensure that they are aligned with human values and norms. This involves creating systems that can adapt to the specific needs of communities and embedding moral values into these systems to guide their actions. By drawing inspiration from human social structures, we can create technology that serves our needs and respects the complex moral landscape in which we live.