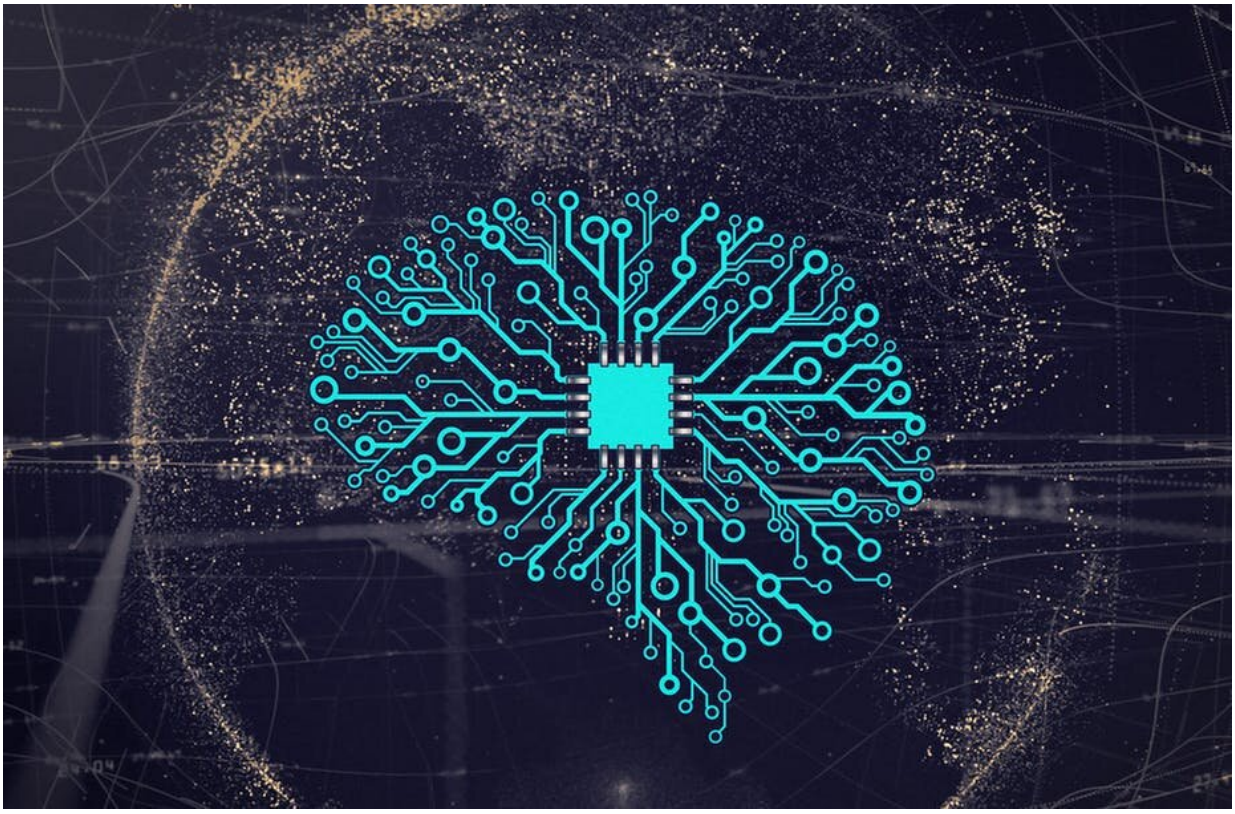
Artificial intelligence and algorithmic irresponsibility: The devil in the machine?

March 17 2021, by Ismael Al-Amoudi



Today, artificial intelligence is deeply imbedded in the systems we use to make decisions. However, the assumptions on which they're built are often completely hidden to us. Credit: mikemacmarketing, CC BY

The classic 1995 crime film *The Usual Suspects* revolves around the police interrogation of Roger "Verbal" Kint, played by Kevin Spacey.

Kint paraphrases Charles Baudelaire, stating that "the greatest trick the Devil ever pulled was convincing the world he didn't exist." The implication is that the Devil is more effective when operating unseen, manipulating and conditioning behavior rather than telling people what to do. In the film's narrative, his role is to cloud judgment and tempt us to abandon our sense of moral responsibility.

In our research, we see parallels between this and the role of artificial intelligence (AI) in the 21st century. Why? AI tempts people to abandon judgment and moral responsibility in just the same way. By removing a range of decisions from our conscious minds, it crowds out judgment from a bewildering array of human activities. Moreover, without a proper understanding of how it does this we cannot circumvent its negative effects.

The role of AI is so widely accepted in 2020 that most people are in essence completely unaware of it. Among other things, today AI algorithms help determine who we date, our medical diagnoses, our investment strategies, and what exam grades we get.

Serious advantages, insidious effects

With widespread access to granular data on human behavior harvested from social media, AI has permeated the key sectors of most developed economies. For tractable problems such as analyzing documents, it usually compares favorably with human alternatives that are slower and more error-prone, leading to enormous efficiency gains and cost reductions for those who adopt it. For more complex problems such as choosing a life-partner, AI's role is more insidious: it frames choices and "nudges" choosers.

It is for these more complex problems that we see substantial risk associated to the rise of AI in decision-making. Every human choice

necessarily involves transforming inputs (relevant information, feelings, etc.) into outputs (decisions). However every choice inevitably also involves a *judgment* – without judgment we might speak of a reaction rather than a choice. The judgmental aspect of choice is what allows humans to attribute responsibility. But as more complex and important choices are made, or at least driven, by AI, the attribution of responsibility becomes more difficult. And there is a risk that both public and private sector actors embrace this erosion of judgment and adopt AI algorithms precisely in order to insulate themselves from blame.

In a recent research paper, we have examined how reliance on AI in health policy may obfuscate important moral discussions and thus "deresponsibilize" actors in the health sector. (See "Anormative black boxes: artificial intelligence and health policy," *Post-Human Institutions and Organizations: Confronting the Matrix*.)

Erosion of judgment and responsibility

Our research's key insights are valid for a wider variety of activities. We argue that the erosion of judgment engendered by AI blurs—or even removes—our sense of responsibility. The reasons are:

- **AI systems operate as black boxes**. We can know the input and the output of an AI system, but it is extraordinarily tricky to trace back how outputs were deduced from inputs. This apparently intractable opacity generates a number of moral problems. A black box can be causally responsible for a decision or action, but cannot explain how it has reached that decision or recommended that action. Even if experts open the black box and analyze the long sequences of calculations that it contains, these cannot be translated into anything resembling a human justification or explanation.

- **Blaming impersonal systems of rules**. Organizational scholars have long studied how bureaucracies can absolve individuals of the worst crimes. Classic texts include Zygmunt Bauman's *Modernity and the Holocaust* and Hannah Arendt's *Eichmann in Jerusalem*. Both were intrigued by how otherwise decent people could participate in atrocities without feeling guilt. This phenomenon was possible because individuals shifted responsibility and blame to impersonal bureaucracies and their leaders. The introduction of AI intensifies this phenomenon because now even leaders can shift responsibility to the AI systems that issued policy recommendations and framed policy choices.
- **Attributing responsibility to artifacts rather than root causes**. AI systems are designed to recognize patterns. But, contrary to human beings, they do not understand the meaning of these patterns. Thus, if most crime in a city is committed by a certain ethnic group, the AI system will quickly identify this correlation. However, it will not consider whether this correlation is an artifact of deeper, more complex, causes. Thus, an AI system can instruct police to discriminate between potential criminals based on skin color, but cannot understand the role played by racism, police brutality and poverty in causing criminal behavior in the first place.
- **Self-fulfilling prophecies that are not blameable on anyone**. Most widely used AIs are fed by historical data. This can work in the case of detecting physiological conditions such as skin cancers. The problem, however, is that AI-classification of *social categories* can operate as a self-fulfilling prophecy in the long run. For instance, researchers on AI-based gender discrimination acknowledge the intractability of algorithms that end up exaggerating, without ever introducing, pre-existing social bias against women, transgendered and non-binary persons.

What can we do?

There is no silver bullet against AI's deresponsibilizing tendencies and it is not our role, as scholars and scientists, to decide when AI-based input should be taken for granted and when it should be contested. This is a decision best left to democratic deliberation. (See "Digital society's techno-totalitarian matrix" in *Post-Human Institutions and Organizations: Confronting the Matrix*.) It is, however, our role to stress that, in the current state of the art, AI-based calculations operate as black boxes that make moral decision-making more, rather than less, difficult.

Provided by The Conversation