Two models for agentic AI

AI ‘agents’ are coming, whether we are ready or not. There can be little doubt that this development will be transformative — for better or worse

AI ‘agents’ are coming, whether we are ready or not. While there is much uncertainty about when AI models will be able to interact autonomously with digital platforms, other AI tools and even humans, there can be little doubt that this development will be transformative — for better or worse. Yet, despite all the commentary (and hype) around agentic AI, many big questions remain unaddressed, the biggest being which type of AI agent the tech industry is seeking to develop?

Different models will have vastly different implications. With an ‘AI as adviser’ approach, AI agents would offer individually calibrated recommendations to human decision-makers, leaving humans always in the driver’s seat. But with an ‘autonomous AI’ model, agents will take the wheel on behalf of humans. That is a distinction with profound and far-reaching implications.

Humans make hundreds of decisions every day, some of which have major consequences for their careers, livelihoods or happiness. Many of these decisions are based on imperfect or incomplete information, determined more by emotions, intuitions, instincts or impulses. As David Hume famously put it, ‘Reason is and ought only to be the slave of the passions.’ Humans may make most decisions without systematic reasoning or due attention to the full implications, but as Hume recognised with the ‘ought’ part of his statement, this isn’t all bad. It is what makes us human. Passion reflects purpose, and it may also play a key role in how we cope with a complex world.

With AI advisers that provide customised, reliable, context-relevant, useful information, many important decisions can be improved, but human motives will remain dominant. But what’s so bad about autonomous AIs making decisions on our behalf? Couldn’t they improve decision-making even further, save time and prevent errors?

A game of chicken

There are several problems with this perspective. First, human agency is critical for human learning and flourishing. The very act of making decisions and contemplating outcomes – even if the inputs and advice come from nonhuman agents – affirms our own sense of agency and purpose. Much of what humans do is not about computation or collecting inputs to decide on an optimal course of action; rather, it is about discovery — an experience that will become increasingly rare if all decisions are delegated to an AI agent.

Moreover, if the tech industry mainly pursues autonomous AI agents, the likelihood of automating more human jobs will increase substantially. Yet, if AI becomes primarily a means of accelerating automation, any hope of widely shared prosperity will be dashed.

Most importantly, there is a fundamental difference between AI agents acting on behalf of humans and humans acting for themselves. Many settings in which humans interact have both cooperative and conflictual elements. Consider the case of one company providing an input to another. If this input is sufficiently valuable to the buyer, a trade between the two companies is mutually beneficial (and typically also benefits society).

A world of autonomous AI agents is likely to usher in many new problems, while eroding many of the gains the technology might have offered.

But for there to be any exchange, the price of the input must be determined through an inherently conflictual process. The higher the price, the more the seller will benefit relative to the buyer. The outcome of such bargaining is often determined by a combination of norms (such as about fairness), institutions (such as contracts that will impose costs if violated) and market forces (such as whether the seller has the option of selling to somebody else). But imagine that the buyer has a reputation for being completely uncompromising — for refusing to accept anything but the lowest feasible price. If there are no other buyers, the seller may be forced to accept the low-ball offer.

Fortunately, in our day-to-day transactions, such uncompromising stances are rare, partly because it pays not to have a bad reputation and, more importantly, because most humans have neither the nerve nor the aspiration to act in such aggressive ways. But now imagine that the buyer has an autonomous AI agent that does not care about human niceties and

possesses nonhuman steely nerves. The AI can be trained always to adopt this uncompromising stance, and the counterparty will have no hope of coaxing it toward a more mutually beneficial outcome. By contrast, in an AI-as-adviser world, the model might still recommend an uncompromising position, but the human would ultimately decide whether to go down that path.

In the near term, then, autonomous agentic AIs may usher in a more unequal world, where only some companies or individuals have access to highly capable, credibly hard-nosed AI models. But even if everyone eventually acquired the same tools, that would not be any better. Our entire society would be subjected to 'war-of-attrition' games in which AI agents push every conflictual situation to the brink of breakdown.

Such confrontations are inherently risky. As in a game of 'chicken' (when two cars accelerate toward each other to see who will swerve away first), it is always possible that neither party will cave. When that happens, both drivers 'win' — and both perish.

An AI that has been trained to win at 'chicken' will never swerve. While AI could be a good adviser to humans – furnishing us with useful, reliable and relevant information in real time – a world of autonomous AI agents is likely to usher in many new problems, while eroding many of the gains the technology might have offered.

© Project Syndicate



Daron Acemoglu
Boston

Daron Acemoglu is Professor of Economics at MIT and co-author of *Why Nations Fail: The Origins of Power, Prosperity and Poverty.*