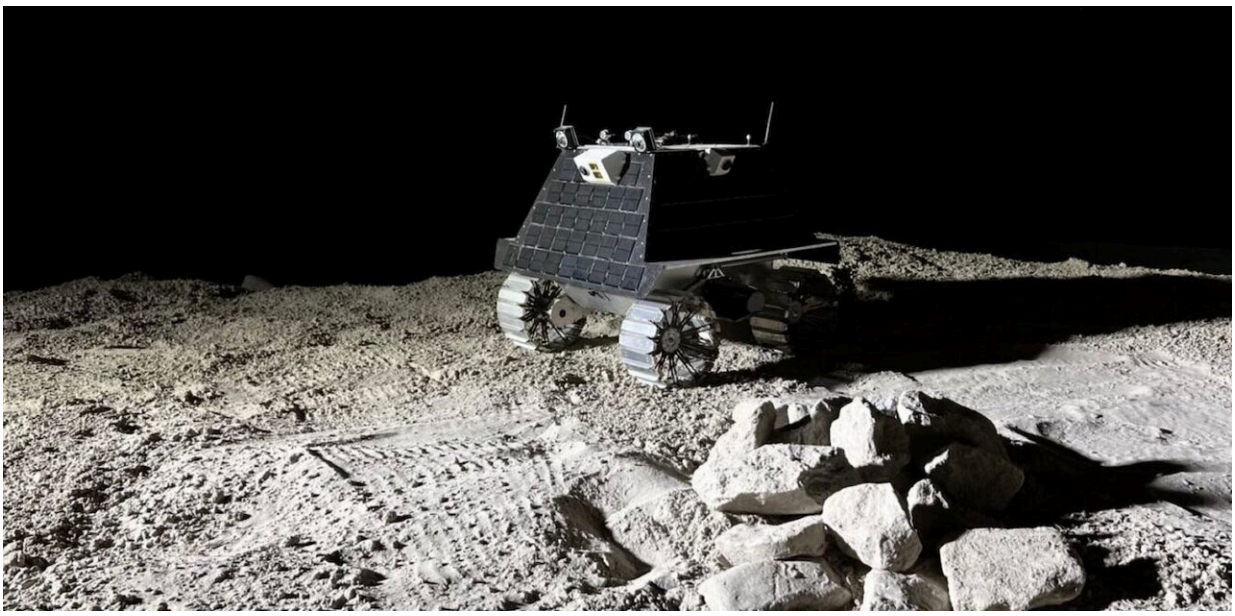
How should a robot explore the moon? A simple question shows the limits of current AI systems

June 28 2023, by Sally Cripps, et al.



Credit: University of Alberta

Rapid progress in artificial intelligence (AI) has spurred some leading voices in the field to call for a research pause, raise the possibility of AI-driven human extinction, and even ask for government regulation. At the heart of their concern is the idea AI might become so powerful we lose control of it.

But have we missed a more fundamental problem?

Ultimately, AI systems should help humans make better, more accurate decisions. Yet even the most impressive and flexible of today's AI tools—such as the large language models behind the likes of ChatGPT—can have the opposite effect.

Why? They have two crucial weaknesses. They do not help decision-makers understand causation or uncertainty. And they create incentives to collect huge amounts of data and may encourage a lax attitude to privacy, legal and ethical questions and risks.

Cause, effect and confidence

ChatGPT and other "foundation models" use an approach called deep learning to trawl through enormous datasets and identify associations between factors contained in that data, such as the patterns of language or links between images and descriptions. Consequently, they are great at interpolating—that is, predicting or filling in the gaps between known values.

Interpolation is not the same as creation. It does not generate knowledge, nor the insights necessary for decision-makers operating in complex environments.

However, these approaches require huge amounts of data. As a result, they encourage organizations to assemble enormous repositories of data—or trawl through existing datasets collected for other purposes. Dealing with "big data" brings considerable risks around security, privacy, legality and ethics.

In low-stakes situations, predictions based on "what the data suggest will happen" can be incredibly useful. But when the stakes are higher, there

are two more questions we need to answer.

The first is about how the world works: "what is driving this outcome?" The second is about our knowledge of the world: "how confident are we about this?"

From big data to useful information

Perhaps surprisingly, AI systems designed to infer causal relationships don't need "big data". Instead, they need *useful information*. The usefulness of the information depends on the question at hand, the decisions we face, and the value we attach to the consequences of those decisions.

To paraphrase the US statistician and writer Nate Silver, the amount of truth is approximately constant irrespective of the volume of data we collect.

So, what is the solution? The process starts with developing AI techniques that tell us what we genuinely don't know, rather than producing variations of existing knowledge.

Why? Because this helps us identify and acquire the minimum amount of valuable information, in a sequence that will enable us to disentangle causes and effects.

A robot on the moon

Such knowledge-building AI systems exist already.

As a simple example, consider a robot sent to the moon to answer the question, "What does the moon's surface look like?"

The robot's designers may give it a prior "belief" about what it will find, along with an indication of how much "confidence" it should have in that belief. The degree of confidence is as important as the belief, because it is a measure of what the robot doesn't know.

The robot lands and faces a decision: which way should it go?

Since the robot's goal is to learn as quickly as possible about the moon's surface, it should go in the direction that maximizes its learning. This can be measured by which new knowledge will reduce the robot's uncertainty about the landscape—or how much it will increase the robot's confidence in its knowledge.

The robot goes to its new location, records observations using its sensors, and updates its belief and associated confidence. In doing so it learns about the moon's surface in the most efficient manner possible.

Robotic systems like this—known as "active SLAM" (Active Simultaneous Localisation and Mapping)—were first proposed more than 20 years ago, and they are still an active area of research. This approach of steadily gathering knowledge and updating understanding is based on a statistical technique called Bayesian optimization.

Mapping unknown landscapes

A decision-maker in government or industry faces more complexity than the robot on the moon, but the thinking is the same. Their jobs involve exploring and mapping unknown social or economic landscapes.

Suppose we wish to develop policies to encourage all children to thrive at school and finish high school. We need a conceptual map of which actions, at what time, and under what conditions, will help to achieve these goals.

Using the robot's principles, we formulate an initial question: "Which intervention(s) will most help children?"

Next, we construct a draft conceptual map using existing knowledge. We also need a measure of our confidence in that knowledge.

Then we develop a model that incorporates different sources of information. These won't be from robotic sensors, but from communities, lived experience, and any useful information from recorded data.

After this, based on the analysis informing the community and stakeholder preferences, we make a decision: "Which actions should be implemented and under which conditions?"

Finally, we discuss, learn, update beliefs and repeat the process.

Learning as we go

This is a "learning as we go" approach. As new information comes to hand, new actions are chosen to maximize some pre-specified criteria.

Where AI can be useful is in identifying what information is most valuable, via algorithms that quantify what we don't know. Automated systems can also gather and store that information at a rate and in places where it may be difficult for humans.

AI systems like this apply what is called a Bayesian decision-theoretic framework. Their models are explainable and transparent, built on explicit assumptions. They are mathematically rigorous and can offer guarantees.

They are designed to estimate causal pathways, to help make the best

intervention at the best time. And they incorporate human values by being co-designed and co-implemented by the communities that are impacted.

We do need to reform our laws and create new rules to guide the use of potentially dangerous AI systems. But it's just as important to choose the right tool for the job in the first place.

Provided by The Conversation