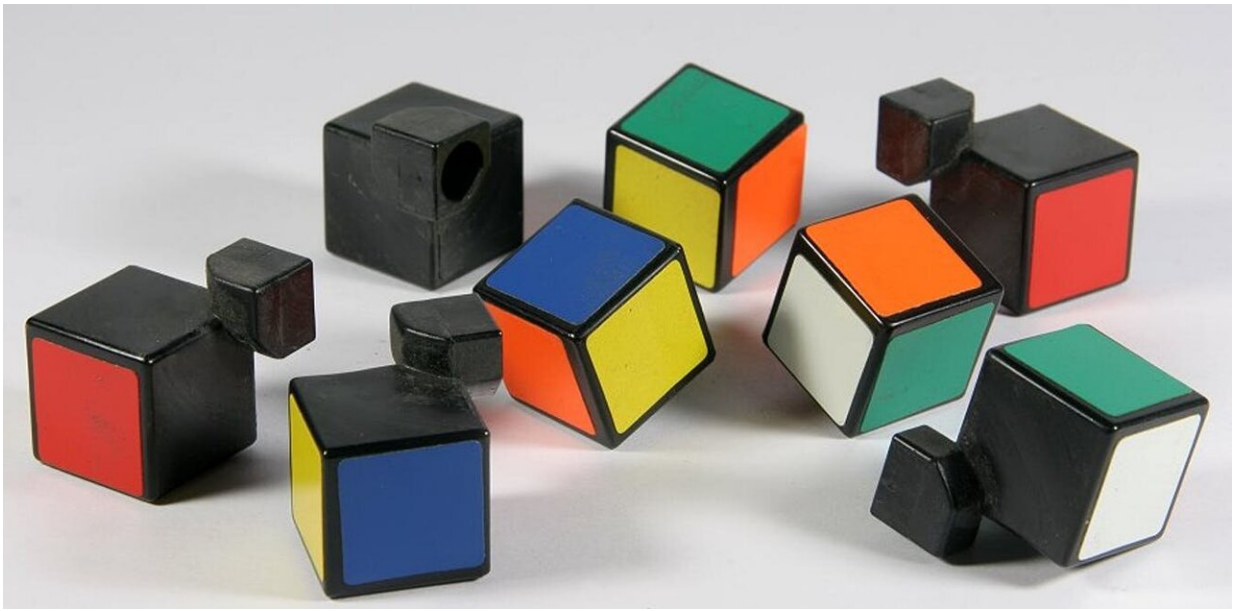
How explainable artificial intelligence can help humans innovate

January 13 2021, by Forest Agostinelli



Understanding how artificial intelligence algorithms solve problems like the Rubik's Cube makes AI more useful. Credit: Roland Frisch via Wikimedia Commons, CC BY-SA

The field of artificial intelligence (AI) has created computers that can drive cars, synthesize chemical compounds, fold proteins and detect high-energy particles at a superhuman level.

However, these AI algorithms cannot explain the thought processes behind their decisions. A computer that masters protein folding and also

tells researchers more about the rules of biology is much more useful than a computer that folds proteins without explanation.

Therefore, AI researchers like me are now turning our efforts toward developing AI algorithms that can explain themselves in a manner that humans can understand. If we can do this, I believe that AI will be able to uncover and teach people new facts about the world that have not yet been discovered, leading to new innovations.

Learning from experience

One field of AI, called reinforcement learning, studies how computers can learn from their own experiences. In reinforcement learning, an AI explores the world, receiving positive or negative feedback based on its actions.

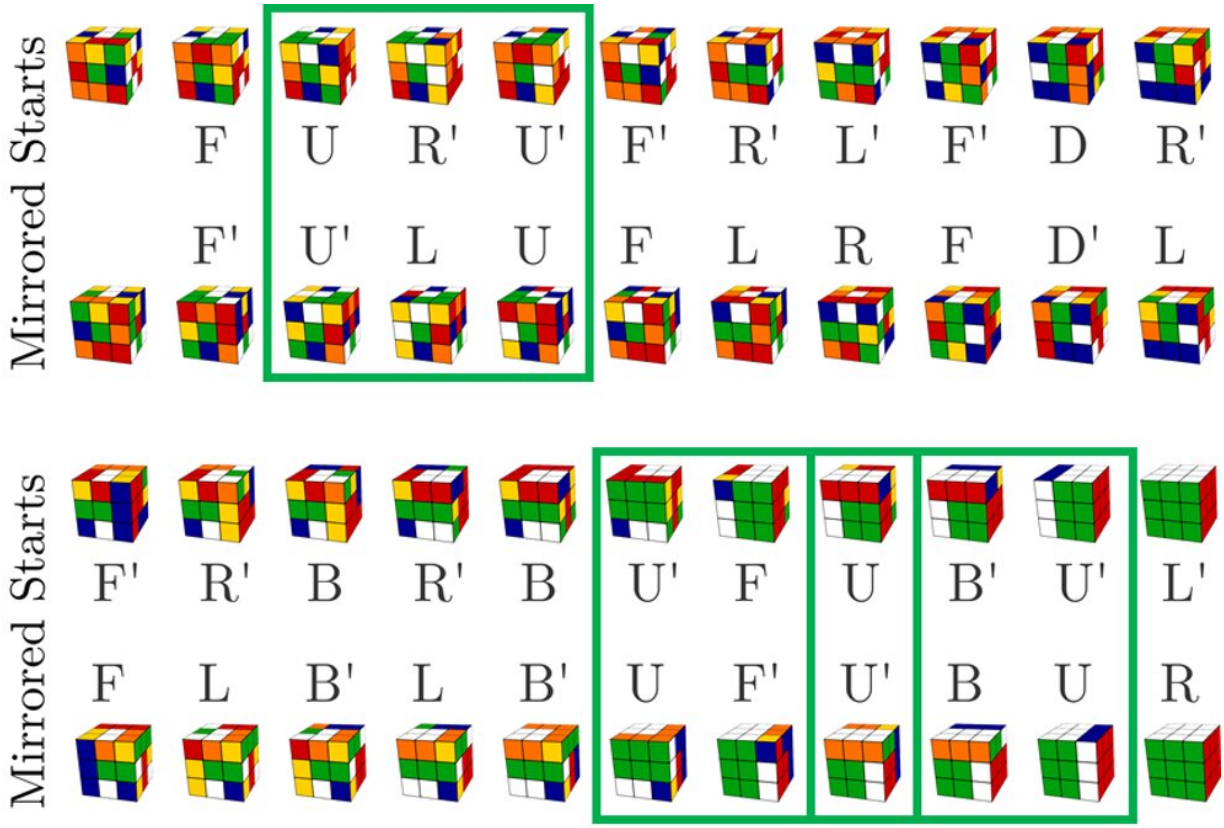
This approach has led to algorithms that have independently learned to play chess at a superhuman level and prove mathematical theorems without any human guidance. In my work as an AI researcher, I use reinforcement learning to create AI algorithms that learn how to solve puzzles such as the Rubik's Cube.

Through reinforcement learning, AIs are independently learning to solve problems that even humans struggle to figure out. This has got me and many other researchers thinking less about what AI can learn and more about what humans can learn from AI. A computer that can solve the Rubik's Cube should be able to teach people how to solve it, too.

Peering into the black box

Unfortunately, the minds of superhuman AIs are currently out of reach to us humans. AIs make terrible teachers and are what we in the

computer science world call "black boxes."



A step-by-step refinement approach can make it easier for humans to understand why AIs do the things they do. Credit: Forest Agostinelli, CC BY-ND

A black-box AI simply spits out solutions without giving reasons for its solutions. Computer scientists have been trying for decades to open this black box, and recent research has shown that many AI algorithms actually do think in ways that are similar to humans. For example, a computer trained to recognize animals will learn about different types of eyes and ears and will put this information together to correctly identify the animal.

The effort to open up the black box is called explainable AI. My research group at the AI Institute at the University of South Carolina is interested in developing explainable AI. To accomplish this, we work heavily with the Rubik's Cube.

The Rubik's Cube is basically a pathfinding problem: Find a path from point A—a scrambled Rubik's Cube—to point B—a solved Rubik's Cube. Other pathfinding problems include navigation, theorem proving and chemical synthesis.

My lab has set up a website where anyone can see how our AI algorithm solves the Rubik's Cube; however, a person would be hard-pressed to learn how to solve the cube from this website. This is because the computer cannot tell you the logic behind its solutions.

Solutions to the Rubik's Cube can be broken down into a few generalized steps—the first step, for example, could be to form a cross while the second step could be to put the corner pieces in place. While the Rubik's Cube itself has over 10 to the 19th power possible combinations, a generalized step-by-step guide is very easy to remember and is applicable in many different scenarios.

Approaching a problem by breaking it down into steps is often the default manner in which people explain things to one another. The Rubik's Cube naturally fits into this step-by-step framework, which gives us the opportunity to open the black box of our algorithm more easily. Creating AI algorithms that have this ability could allow people to collaborate with AI and break down a wide variety of complex problems into easy-to-understand steps.

Collaboration leads to innovation

Our process starts with using one's own intuition to define a step-by-step

plan thought to potentially solve a complex problem. The algorithm then looks at each individual step and gives feedback about which steps are possible, which are impossible and ways the plan could be improved. The human then refines the initial plan using the advice from the AI, and the process repeats until the problem is solved. The hope is that the person and the AI will eventually converge to a kind of mutual understanding.

Currently, our algorithm is able to consider a human plan for solving the Rubik's Cube, suggest improvements to the plan, recognize plans that do not work and find alternatives that do. In doing so, it gives feedback that leads to a step-by-step plan for solving the Rubik's Cube that a person can understand. Our team's next step is to build an intuitive interface that will allow our algorithm to teach people how to solve the Rubik's Cube. Our hope is to generalize this approach to a wide range of pathfinding problems.

People are intuitive in a way unmatched by any AI, but machines are far better in their computational power and algorithmic rigor. This back and forth between man and machine utilizes the strengths from both. I believe this type of collaboration will shed light on previously unsolved problems in everything from chemistry to mathematics, leading to new solutions, intuitions and innovations that may have, otherwise, been out of reach.

This article is republished from The Conversation under a Creative Commons license. Read the original article.

Provided by The Conversation

Citation: How explainable artificial intelligence can help humans innovate (2021, January 13) retrieved 24 February 2024 from https://techxplore.com/news/2021-01-artificial-intelligence-

humans.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.