tions ongoing in their cores, which can last for hundreds of millions to many billions of years, depending on their mass. When such a star exhausts the hydrogen fuel in its core, it expands enormously, shatters any close-enough planet, and becomes a white dwarf. Thereafter, what remains of the planetary system may move close enough to the star to become subject to collisions and to strong tidal forces, grinding the remaining planetary cores (*3*). This leaves behind a shroud of rocky debris of various sizes, ranging from micrometer-sized dust particles to kilometer-sized bodies (*4*, *5*). In some cases, because of the high temperature and strong irradiation present in proximity of a white dwarf, these rocks release metal-rich gas, giving rise to a disk of gas and debris surrounding the white dwarf. The presence of circumstellar gas is indicated by metal emission lines in the stellar spectrum (*6*).

Only a few white dwarfs are known to host a gas disk, and the rocky body detected by Manser *et al.* orbits one of those. Because of the chaotic motion present in the disk surrounding 20 to 25% of white dwarfs (also known as "polluted" white dwarfs), there is a continuous infall of rocky, planetary material onto the stellar surface, which reveals itself through the presence of metal absorption lines in the stellar spectrum (*7*, *8*). This accretion of rocky material is continuous because the strong stellar gravity brings any metal lying on the surface into the inner layers within a very short time scale (*7*, *9*). Therefore, by analyzing the metal absorption and emission lines in the spectrum of a polluted white dwarf, it is possible to identify the composition of the circumstellar gas and/or rocks forming the disk (*2*). The study of Manser *et al.* also concluded that the density of the planetesimal should be between 7.7 and 39 g/cm^3, which is compatible with that of pure iron and of Earth's core. It is therefore plausible that the planetesimal is the remnant core of a shattered planet.

> ***"...by analyzing...the spectrum of a polluted white dwarf, it is possible to identify the composition of the circumstellar gas and/or rocks forming the disk."***

Theoretical models of the orbital evolution of planetary systems indicate that possibly large (a few to hundreds of kilometers in diameter), rocky bodies might survive the last stages of stellar evolution toward the white dwarf phase (*10*, *11*). Furthermore, the existence of numerous polluted white dwarfs indicates that planetesimals indeed orbit around these stars. However, planetesimals orbiting white dwarfs have been directly found in just one case using the Kepler space telescope and the transit method (*12*), despite the large number of polluted white dwarfs discovered to date, the fact that white dwarfs are the descendant of almost all planet hosts known to date, and that their small size facilitates the detection of transiting bodies.

The method of Manser *et al.* has revealed the presence of planetesimals without the need for the particular orbital geometry that is required by the transit method. It could therefore be used to identify the presence of planetesimals orbiting other polluted white dwarfs and advance the study of the planetary systems evolution. Furthermore, because planetesimals orbiting white dwarfs are believed to be the remnant cores of shattered planets, studying the spectra of polluted white dwarfs known to be surrounded by planetesimals enables one to gain information about the chemical composition and metal abundances of the infalling material—that is, planetary cores (*13*). This kind of characterization is not possible for bodies in the solar system, including Earth.

Because of their small size, white dwarfs are faint. The discovery of Manser *et al.* required observations conducted with the 10.4-m Gran Telescopio Canarias in La Palma, Spain, which is one of the largest in the world. Future similar discoveries will therefore require high-efficiency instruments and large telescopes. The range of extremely large telescopes in Chile and Hawaii, currently under construction or planned, will have primary mirrors that are 30 to 40 m in diameter. This should be the ideal platform for finding more planetesimals orbiting white dwarfs and exploring the innermost regions of planets. ■

REFERENCES AND NOTES

1. C. J. Manser *et al.*, *Science* **364**, 66 (2019).
2. M. A. Hollands, B. T. Gänsicke, D. Koester, *Mon. Not. R. Astron. Soc.* **477**, 93 (2018).
3. J. Farihi *et al.*, *Mon. Not. R. Astron. Soc.* **481**, 2601 (2018).
4. J. C. Brown, D. Veras, B. T. Gänsicke, *Mon. Not. R. Astron. Soc.* **468**, 1575 (2017).
5. A. J. Mustill, E. Villaver, D. Veras, B. T. Gänsicke, A. Bonsor, *Mon. Not. R. Astron. Soc.* **476**, 3939 (2018).
6. B. T. Gänsicke, T. R. Marsh, J. Southworth, A. Rebassa-Mansergas, *Science* **314**, 1908 (2006).
7. B. Zuckerman, C. Melis, B. Klein, D. Koester, M. Jura, *Astrophys. J.* **722**, 725 (2010).
8. D. Koester, B. T. Gänsicke, J. Farihi, *Astron. Astrophys.* **566**, A34 (2014).
9. J. Farihi *et al.*, *Mon. Not. R. Astron. Soc.* **463**, 3186 (2016).
10. D. Veras, Z. M. Leinhardt, A. Bonsor, B. T. Gänsicke, *Mon. Not. R. Astron. Soc.* **445**, 2244 (2014).
11. D. Veras, B. T. Gänsicke, *Mon. Not. R. Astron. Soc.* **447**, 1049 (2015).
12. A. Vanderburg *et al.*, *Nature* **526**, 546 (2015).
13. D. J. Wilson, B. T. Gänsicke, J. Farihi, D. Koester, *Mon. Not. R. Astron. Soc.* **459**, 3282 (2016).

10.1126/science.aax0051

ARTIFICIAL INTELLIGENCE

In defense of the black box

Black box algorithms can be useful in science and engineering

By **Elizabeth A. Holm**

The science fiction writer Douglas Adams imagined the greatest computer ever built, Deep Thought, programmed to answer the deepest question ever asked: the Great Question of Life, the Universe, and Everything. After 7.5 million years of processing, Deep Thought revealed its answer: Forty-two (*1*). As artificial intelligence (AI) systems enter every sector of human endeavor—including science, engineering, and health—humanity is confronted by the same conundrum that Adams encapsulated so succinctly: What good is knowing the answer when it is unclear why it is the answer? What good is a black box?

In an informal survey of my colleagues in the physical sciences and engineering, the top reason for not using AI methods such as deep learning, voiced by a substantial majority, was that they did not know how to interpret the results. This is an important objection, with implications that range from practical to ethical to legal (*2*). The goal of scientists and the responsibility of engineers is not just to predict what happens but to understand why it happens. Both an engineer and an AI system may learn to predict whether a bridge will collapse. But only the engineer can explain that decision in terms of physical models that can be communicated to and evaluated by others. Whose bridge would you rather cross?

Scientists and engineers are not alone in their skepticism of black box answers. The European Union General Data Protection Regulation (GDPR), introduced in 2018, guarantees subjects "meaningful information about the logic involved" in automatic decision-making based on their personal data (*3*). The legal interpretation of this regulation is under debate, but the mistrust of inexplicable systems is evident in the statute.

Department of Materials Science and Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA. Email: eaholm@andrew.cmu.edu

In this general atmosphere of suspicion, it is not surprising that AI researchers focus less on defending black box systems and more on understanding how they make decisions, termed the interpretability problem (*4*). In fact, this is one of the grand challenges in current computer science. But this blanket rejection of black box methods may be hasty. In reality, scientists and engineers—like all humans—base many decisions on judgment and experience, which are the outcomes of their own "deep learning" (*5*). As a result, neuroscience struggles with the same interpretability challenge as computer science (*6*). Yet, we routinely accept human conclusions without fully understanding their origin. In this context, it seems reasonable to consider whether black box answers generated by AI systems have a similarly useful role and, if so, when we should apply them (see the figure).

The first and most obvious case for using a black box is when the cost of a wrong answer is low relative to the value of a correct answer. Targeted advertising is the canonical example. From the vendor's point of view, the cost of posting an unwanted ad is small, whereas the benefit of a successful ad is potentially large (*7*). In my own field of materials science, image segmentation—the task of categorizing the pixels in a picture—typically involves a human manually outlining the objects of interest in an image of the complex, internal substructure of a material. This is a costly process, so much so that Ph.D. theses and industrial quality-control systems are designed to require as little image segmentation as possible. An AI system can be trained to do this job with high, but not perfect, fidelity (*8*). Perfection is not, however, necessary to make this system useful because the cost of a few disputed pixels is low compared with saving the time and sanity of belabored graduate students.

The second case for the black box is equally obvious but more fraught. A black box can and should be used when it produces the best results. For example, in reading standard field-of-view medical images, trained AI systems enhance the performance of human radiologists at detecting cancers (*9*). Although the cost of a wrong answer, whether a false negative or a false positive, may be high, the black box offers the best solution that is currently available. Of course, letting AIs read mammograms is not controversial, in part because a human doctor checks the answer. Letting AIs drive cars is more contentious because the black box necessarily makes life-or-death decisions without an opportunity for human intervention. That said, self-driving vehicles eventually will be safer than those piloted by humans; they will produce the best results with respect to traffic injuries and fatalities. When that crossover point occurs can be determined with appropriate objective metrics (*10*), but the societal choice whether to cede human agency to the AI drivers will inevitably involve decisions based on subjective factors, including how to apply human values of ethics, fairness, and accountability to nonhuman entities.

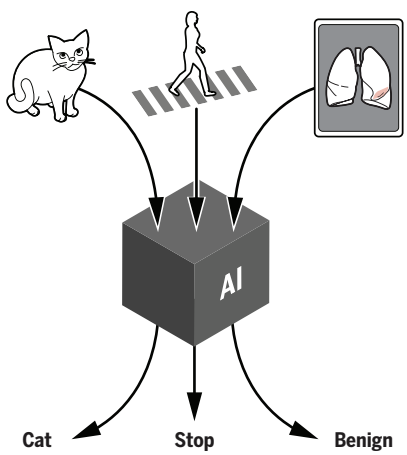
These arguments should not be interpreted as a free license to apply black box methods. The two use cases above presume an ideal black box operated by a user who can compute costs and define best results unambiguously. Both assumptions are subject to pitfalls. AI systems may suffer from a host of shortcomings, including biases, inapplicability outside of the training domain, and brittleness (the tendency to be easily fooled) (*11*). Moreover, evaluating costs and best outcomes is a complex and subjective exercise in balancing economic, individual, societal, and ethical considerations. Worse, these factors can compound: A biased model may entail hidden costs, both from objectively wrong predictions and subjectively measured unfairness. A brittle model may have blind spots that cause spectacularly bad decisions. As with any decision-making system, the black box must be used with knowledge, judgment, and responsibility.

Although AI thought processes can be limited, biased, or outright wrong, they are also different from human thought processes in ways that can reveal new connections and approaches. This brings us to the third case for black box systems: as tools to inspire and guide human inquiry. For example, in a groundbreaking medical imaging study, scientists trained a deep learning system to diagnose diabetic retinopathy—a diabetes complication that affects the eyes—from retinal images. They achieved performance that met or surpassed a committee of ophthalmological experts (*12*, *13*). More surprisingly, the system could accurately identify a number of other characteristics that are not normally assessed with retinal images, including cardiological risk factors, age, and gender (*14*). No one had previously noticed gender-based differences in human retinas, so the black box observation inspired researchers to investigate how and why male and female retinas differ. Pursuing those questions took them away from the black box in favor of interpretable artificial and human intelligence.

Which returns us to the problem with Deep Thought's answer. We cannot use black box AI to find causation, systematization, or understanding. A black box cannot tell us how or why a bridge collapses or what is the great question of Life, the Universe, and Everything. At least for now, these questions remain the purview of human intelligence and the broad and growing field of interpretable AI. In the meantime, however, it is worth accepting the black box on its own terms. Black box methods can contribute substantively and productively to science, technology, engineering, and math to provide value, optimize results, and spark inspiration. ■

When a black box is valuable

By definition, humans cannot assess how a black box algorithm arrives at a particular answer. However, black box methods can still provide value when they produce the best results, when the cost of a wrong answer is low, or when they inspire new ideas.



GRAPHIC: N. DESAI/*SCIENCE*

REFERENCES AND NOTES

1. D. Adams, *Hitchhiker's Guide to the Galaxy* (Harmony Books, 1980).
2. D. S. Char, N. H. Shah, D. Magnus, *N. Engl. J. Med.* **378**, 981 (2018).
3. Regulation (EU) 2016/679 of the European Parliament and of the Council, 27 April 2016; https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN.
4. F. Doshi-Velez, B. Kim, arXiv:1702.08608v2 [stat.ML] (2017).
5. R. E. Nisbett, T. D. Wilson, *Psychol. Rev.* **84**, 231 (1977).
6. J. I. Gold, M. N. Shadlen, *Annu. Rev. Neurosci.* **30**, 535 (2007).
7. D. S. Evans, *J. Econ. Perspect.* **23**, 37 (2009).
8. B. L. DeCost, B. Lei, T. Francis, E. A. Holm, *Microsc. Microanal.* **25**, 21 (2019).
9. A. Jalalian *et al.*, *Clin. Imaging* **37**, 420 (2013).
10. N. Kalra, S. M. Paddock, *Transport. Res. A* **94**, 182 (2016).
11. G. Marcus, arXiv:1801.00631 [cs.AI] (2018).
12. V. Gulshan *et al.*, *JAMA* **316**, 2402 (2016).
13. J. Krause *et al.*, *Ophthalmology* **125**, 1264 (2018).
14. R. Poplin *et al.*, *Nat. Biomed. Eng.* **2**, 158 (2018).

ACKNOWLEDGMENTS

This work has been funded by U.S. National Science Foundation grants DMR-1507830 and CMMI-1826218 and by the Carnegie Mellon Manufacturing Futures Initiative through a grant from the Richard King Mellon Foundation.

10.1126/science.aax0162

In defense of the black box

Elizabeth A. Holm

Science **364** (6435), 26-27.
DOI: 10.1126/science.aax0162