Artificial Intelligence and Black-Box Medical Decisions: *Accuracy versus Explainability*

BY ALEX JOHN LONDON

Although decision-making algorithms are not new to medicine, the availability of vast stores of medical data, gains in computing power, and breakthroughs in machine learning are accelerating the pace of their development, expanding the range of questions they can address, and increasing their predictive power. In many cases, however, the most powerful machine learning techniques purchase diagnostic or predictive accuracy at the expense of our ability to access "the knowledge within the machine."[1] Without an explanation in terms of reasons or a rationale for particular decisions in individual cases, some commentators regard ceding medical decision-making to black box systems as contravening the profound moral responsibilities of clinicians. As William Swartout puts it, when a physician consults an expert, "[t]he physician may question whether some factor was considered or what effect a particular finding had on the final outcome and the expert is expected to be able to justify his answer and show that sound medical principles and knowledge were used to obtain it. . . . In addition to providing diagnoses or prescriptions, a consultant program must be able to explain what it is doing and justify why it is doing it."[2] To the extent that deep learning systems cannot explain their findings, some have questioned whether medical systems should avoid such approaches and "sacrifice predictive power in favor of simplicity of a model."[3]

As far back as the ancient Greeks, trust has been connected to the ability to explain expert recommendations. We expect that experts can marshal well-developed causal knowledge to explain their actions or recommendations, a feat that is a reality in some modern scientific domains. Against that background expectation, the most powerful machine learning techniques seem woefully incomplete because they are atheoretical, associationist, and opaque. A major problem with this view about the importance of explanation, I argue below, is that empirical findings in medicine often have better epistemic footing than the theories that might explain them and that atheoretical, associationist, and opaque decisions are more common in medicine than critics realize. Moreover, as Aristotle noted over two millennia ago, when our knowledge of causal systems is incomplete and precarious—as it often is in medicine—the ability to explain how results are produced can be less important than the ability to produce such results and empirically verify their accuracy. I conclude with some reasons that a blanket requirement that machine learning systems in medicine be explainable or interpretable is unfounded and potentially harmful.

Justification, Explanation, and Causation

Trust in experts is often grounded in their ability to produce certain results and to justify their actions. As a result, it is sometimes claimed that trust in computational decision-makers must be grounded in more than predictive or diagnostic accuracy. It also requires the ability to justify their recommendations. As Swartout notes, "By justifications, we mean explanations that tell why an expert system's actions are reasonable in terms of principles of the domain—the reasoning behind the system."[4] Explanations of this form require the system, or the expert who relies on it, to reveal how a finding or a decision is grounded in two kinds of knowledge: a "domain model" in which causal relationships in the domain are captured and "domain principles" that lay

Alex John London, "Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability," *Hastings Center Report* 49, no. 1 (2019): 15-21. DOI: 10.1002/hast.973

out the "how to" knowledge or the dynamics of the domain in question.[5]

These requirements seem reasonable, in part, because they have a long intellectual pedigree. Already in the moral thinking of the ancient Greeks there is recognition of different forms of knowledge, suited to different spheres, acquired in different ways, reliable under different circumstances, and amenable to different demands for explanation and justification. In the practical sphere, a techne (a productive science) is concerned with bringing into existence particular effects in a specific domain—intervening in the causal nexus in order to make a particular kind of house, to produce shoes for a particular kind of horse, or to effect a siege against a particular type of stronghold. However, these effects can often be brought about through other means. For example, Aristotle says that empirics—people with a lot of experience in a particular domain but who lack a theory to explain their success—can often achieve better results than people who know only theory.[6] With experience, empirics gain know-how, but they lack an account or an explanation for why their recommendations work. What sets apart techne as a form of productive knowledge is that it includes a theory whose general principles explain why certain actions are correct in particular circumstances.[7] Whereas the empiric knows only to prescribe chicken to preserve health, and the mere theorist knows only that lean meats make people healthy, the science of medicine combines the knowledge of the particular with knowledge of the universal—the reason to prescribe chicken is because it is a lean meat and eating lean meats is a cause of health.

The idea that experts should be able to justify their actions by marshaling knowledge of causal relationships in their domain of expertise also has a long intellectual history. For Aristotle, explanations are logical arguments in which the particular to be explained is subsumed under a more general set of claims that clarify the causal factors responsible for generating the particular.[8] Although our understanding of causality has developed since the time of Aristotle, the idea that an explanation should have something like this form persists.[9] When explanations involve laws that track causal relationships, true explanations provide insight into how a domain works and, through that insight, enhance our ability to more effectively intervene in that system, where intervention is possible.

Another reason to expect computational systems to be able to marshal causal knowledge and provide explanations of this form is that this appears to be a mark of expertise in disciplines such as structural engineering. Building a bridge across a span requires a range of decisions. Expert structural engineers make such decisions by marshaling "domain models" and "domain principles" of the sort that commentators like Swartout expect experts to possess. They know what factors affect the success of a bridge, such as properties of the location, features of materials to consider, and the tolerances of various designs and the stresses of various uses. They also know how to assign values to these variables in particular cases and to simulate how particular structures will behave under expected loads and stresses of a particular setting to within practically relevant margins of error.[10] Detailed mathematical models of key causal relations enable structural engineers to make design decisions that incorporate stakeholder values, such as aesthetics and cost, into the construction of reliable structures. Moreover, they can explain particular decisions by elaborating the functional and causal requirements that constrain or determine various choices, thereby helping nonexperts understand why certain decisions were made or why some constraints are negotiable while others are not. Together, this causal knowledge and the explanations it supports can also guide interventions to improve a structure's integrity.

The classical model of the techne and its modern instantiations in areas like structural engineering thus present a model of decision-making that is highly rational and in which decisions reflect causal knowledge of a domain that can be expressed in terms that are, at least in principle, accessible to nonexperts. These explanations thus help to foster social trust by expanding the ability of other stakeholders to understand what is at stake in various decisions. This fosters accountability, since understanding why a decision was made enables stakeholders to evaluate its merits and hold experts accountable for avoidable error. It also fosters autonomy in the form of nondomination,[11] to the extent that explanations help stakeholders see why expert decisions are not arbitrary and do not amount to abuse of professional authority.

The Black Box of Deep Learning

Against this background, many of the properties of the most powerful machine learning systems appear suspect. For example, deep learning systems are theory agnostic in the sense that their designers do not program into them a model that reflects their understanding of the causal structure of the problem to be solved. Rather, programmers construct an architecture that "learns" a model from a large set of data. This architecture contains layers of connected nodes, like neurons in a brain, that activate when they detect particular features in input data. These systems are "deep" learners in that they contain many nested layers of such nodes. In most cases, these systems learn when data whose classification is already established (for example, images of retinas that display or lack diabetic retinopathy) are fed into the system. As instances accumulate, weights on the nodes in the network are automatically adjusted to construct the mathematical model that most accurately maps inputs (such as images of retinas[12] or patient medical records[13]) to the correct output labels. The systems classify images as displaying diabetic retinopathy or not, or assign a probability for a medical event, such as suicide or readmission, to a medical record. After the training phase, the sensitivity, specificity, and recall of such systems can then be tested by inputting a second set of data whose classification is already known and then comparing output classifications to the "ground truth." Deep learning systems can be trained on millions of inputs, and their resulting predictions can be highly accurate.

The opacity, independence from an explicit domain model, and lack of causal insight associated with some powerful machine learning approaches are not radically different from routine aspects of medical decision-making.

Despite this accuracy, deep learning systems can be black boxes. Although their designers understand the architecture of these systems and the process by which they generate the models they use for classification, the models themselves can be inscrutable to humans. Even when techniques are used to identify features or a set of features to which a model gives significant weight in evaluating a particular case, the relationships between those features and the output classification can be both indirect and fragile. A small permutation in a seemingly unrelated aspect of the data can result in a significantly different weighting of features. Moreover, different initial settings can result in the construction of different models.[14]

Despite the overwhelming attention paid to the fact that deep learning systems are unsuited to helping human users understand the phenomenon in question, a far more significant limitation is that they may not directly track causal relationships in the world. Even when users limit the data fed into the system to variables believed to be causally relevant to the decision at hand, the resulting model only reflects regularities in data. How these associations relate to underlying causal relationships is unknown. Even if we can learn that a system associates having "cocaine test: negative" in a patient's electronic medical record with a higher likelihood of readmission,[15] this knowledge doesn't reveal what a negative test indicates or how this could be causally related to whatever causes readmission. As a result, understanding that a system uses a negative cocaine test as a predictor of readmission doesn't increase our ability to more effectively intervene in the system being modeled.

In contrast to the logical and accessible decision-making embodied in the classical techne, machine learning systems stoke fears of unaccountability and domination by systems that arbitrarily restrict stakeholder autonomy and represent a conduit for experts to covertly impose arbitrary preferences on stakeholders.

Uncertainty and Incompleteness of Medical Knowledge

Given the explanatory power of productive sciences like structural engineering and the long history of regarding medicine as a paradigmatic example of a techne, it seems reasonable to expect medical experts to live up to the same standards as the structural engineer. The problem with this view is that the explanatory power of fields like structural engineering derives from the degree of comparative completeness with which the relevant causal systems are known. Although medicine is one of the oldest productive sciences, its knowledge of underlying causal systems is in its infancy; the pathophysiology of disease is often uncertain, and the mechanisms through which interventions work is either not known or not well understood. As a result, decisions that are atheoretic, associationist, and opaque are commonplace in medicine.

Medicine is a domain in which the ability to intervene effectively in the world by exploiting particular causal relationships often derives from experience and precedes our ability to understand why interventions work—our ability to accurately model causal relationships in a larger portion of the systems in which we intervene. Just as Aristotle's empiric succeeds in promoting health by prescribing chicken for a healthy diet, even without knowing why chicken is a healthy food, modern clinicians prescribed aspirin as an analgesic for nearly a century without understanding the mechanism through which it works. Lithium has been used as a mood stabilizer for half a century, yet why it works remains uncertain. Large parts of medical practice frequently reflect a mixture of empirical findings and inherited clinical culture. In these cases, even efficacious recommendations of experts can be atheoretic in this sense: they reflect experience of benefit without enough knowledge of the underlying causal system to explain how the benefits are brought about.

Randomized clinical trials (RCTs) can establish causal relationships between interventions and measured end points. But the relationship between those end points and the theories that motivate intervention development and guide deployment in practice is more tenuous. The hypothesis that amyloid plaques in the brain are part of the disease process of Alzheimer's disease has motivated a decade-long search for neuroprotective interventions that disrupt the amyloid production system. The repeated failure of these efforts may reflect the falsity of the underlying theory or merely the practical difficulty of effectively intervening in the amyloid

system.[16] As a result, the practical findings from rigorous empirical testing are frequently more reliable and reflective of causal relationships than the theoretical claims that purport to ground and explain them.

Medicine is thus a sphere where current theories of disease pathophysiology or drug mechanism are often of unknown or uncertain value. Since animal and in vitro models are unreliable predictors of effects in humans, specific hypotheses generated by these theories are subjected to testing during the process of evaluating the interventions that they support and motivate.[17] Although the ambition of contemporary drug development is to leverage expanding knowledge about these factors to produce a more analytical and logical development process, roughly nine of ten drugs that enter development are never approved for any indication—and half of the drugs that enter phase III testing fail.[18] Hidden within this summary statistic is the fact that in some areas (for example, developing neuroprotective treatments against Parkinson's or Alzheimer's disease), nothing that we try has worked. Despite widespread expectations that megadoses of vitamins will have therapeutic or preventative effects in indications from cancer to multiple sclerosis, trials routinely demonstrate no clinical value. In fact, the Carotene and Retinol Efficacy Trial was stopped early after it was clear that participants at high risk of lung cancer who received high doses of beta-carotene and retinyl palmitate had a *higher* incidence of cancer and a higher mortality rate than participants in the control arm.[19]

Although Aristotle thought that techne represented a paradigm of productive knowledge, he also understood that not all branches of decision-making were as well understood as others. For this reason, he warned that we must not "demand in all matters alike an explanation of the reason why things are what they are; in some cases it is enough if the fact that they are so is satisfactorily established."[20] Although we can explain why an arch can bear a particular load, we may not be able to explain why a drug stabilizes mood or eases pain. In both practical science and practical wisdom, Aristotle is explicit that, where we cannot have knowledge of both particular facts and the general principles that explain them, knowledge of the particulars is more important because it is more critical to success in action.[21]

In medicine, the overreliance on theories that explain why something might be the case has sometimes made it more difficult to validate the empirical claims derived from such theories, with disastrous effects. The long medical preference for radical mastectomy over less aggressive alternatives was driven by the pathophysiological theory that removing as much tissue from the breast as possible would reduce the probability of cancer recurrence. Only after a series of clinical trials was this theory shown to be false. The same is true for the theory of drug action that drove the use of high-dose chemotherapy with autologous bone marrow transplant as a treatment for end-stage breast cancer. In such cases, the overreliance on plausible theoretical explanations lead to treatment practices that harmed patients and consumed scarce resources precisely because key causal claims in those theories were false.

Even if the efficacy of a particular intervention for a given indication has been established in large RCTs, patients in the clinic often differ from clinical trial populations. Clinicians, therefore, frequently make judgments about how comorbidities, gender, ethnicity, age, or other factors might affect intervention efficacy and toxicity that go beyond validated medical evidence.[22] Treatments are delivered on the basis of explicit or implicit associations between a network of clinical characteristics. In these cases, it may not be clear what information clinicians draw on to make these judgments, whether the implicit or explicit models that support their judgments are valid or accurate, or whether equally qualified clinicians would arrive at the same conclusions in the face of the same data. Certainly, we should try to generate reliable clinical evidence that can illuminate and guide these decisions. But this kind of uncertainty is a routine part of clinical practice, and the clinical judgment that it involves relies on an associationist model encoded in the neural network in the clinician's head that is opaque and often inaccessible to others.

As counterintuitive and unappealing as it may be, the opacity, independence from an explicit domain model, and lack of causal insight associated with some of the most powerful machine learning approaches are not radically different from routine aspects of medical decision-making. Our causal knowledge is often fragmentary, and uncertainty is the rule rather than the exception. In such cases, careful empirical validation of an intervention's practical merits is the most important task. When the demand for explanations of how interventions work is elevated above careful, empirical validation, patients suffer, resources are wasted, and progress is delayed.

Responsible Medical Decision-Making

If the goal is to secure trust among stakeholders, then the accuracy of a system relative to viable alternatives must be a central concern. One advantage of explicit computational systems over the neural networks inside the heads of expert clinicians is that the reliability and accuracy of the former can be readily evaluated and incrementally improved. In high-volume contexts, such as diagnostic medical imaging, the use of tests that are less sensitive (that is, less likely to detect true cases of a condition), less specific (less likely to exclude only false cases), or less precise (with less likelihood that a positive test result correlates with having the condition) than available alternatives can result in avoidable morbidity and mortality on the part of patients. Any preference for less accurate models—whether computational systems or human decision-makers—carries risks to patient health and welfare.[23] Without concrete assurance that these risks are offset by the expectation of additional benefits to patients, a blanket preference for simpler models is simply a lethal prejudice.

It might be objected that explainability is too demanding a requirement since even simple associationist models are not capable of tracking causal relationships.[24] Nevertheless, defects in the data used by deep learning systems to construct

In medicine, the ability to intervene effectively in the world by exploiting causal relationships often derives from experience and precedes clinicians' ability to understand why interventions work.

decision models—such as biases stemming from the over- or underrepresentation of particular classes of individuals—can be inherited by these systems.[25] Without insight into how the models work, critics worry that the models may incorporate biases that are harmful enough to offset marginal gains in predictive power. In order to ward off such possibilities, critics hold that machine learning systems must at least be interpretable to humans.

In a popular example, Rich Caruana and colleagues report that, although a neural net was more accurate than alternatives at diagnosing the probability of death from pneumonia, it ranked asthmatic patients as having a lower probability than the general population.[26] This finding is "counterintuitive" because patients with a history of asthma are typically admitted directly into the intensive care unit (ICU) for aggressive medical care; it is the added care that gives them a lower probability of death. Without such aggressive care, asthmatic patients have a higher probability of death from pneumonia. Their score in the system is seen as misleading because it doesn't reflect patients' underlying medical need. This prompted Caruana et al. to prefer less accurate but more transparent models in which they could adjust the weight assigned to "asthmatic" to reflect current medical knowledge.

It is important to point out, however, that the ascription of bias in this example presupposes that the goal of the decision model is to optimize the allocation of medical resources against a baseline risk of death that is independent of current medical practice. But insofar as the training data reflect the probability of death given standard medical practice, this is probably an inappropriate expectation for many patients, not just for asthmatics. Everyone's outcomes reflect the effects of a range of practices not captured in the data. But patients with different medical histories or comorbidities are likely to receive different levels of care. A data set that reflects patient outcomes and lacks a comprehensive and granular representation of patient characteristics and treatment practices would probably not accurately reflect probability of death prior to any medical intervention. If given more comprehensive information about treatments administered to individual patients, even a simple system would learn that, without ICU admission, asthma puts a patient at high probability of death.

In contrast, if the goal is to identify patients most at risk of dying *given standard practice*, then systems that rank asthmatics at lower risk are not biased. Rather, the system is actuarially correct—patients with asthma *who receive aggressive medical intervention* have a lower probability of death than some nonasthmatic patients who likely receive less aggressive medical care. Such a system could be used to identify classes of patients who might benefit from care additional to, or more aggressive than, standard practice. However, how to improve the outcomes of different classes of patients is a distinct, causal question that we should not expect this system to answer. Rather than illustrating the need for interpretability, this example illustrates the importance of understanding the kind of judgments that a data set is likely to be able to support, clearly validating the accuracy of those specific decisions on real-world data, and then restricting the use of associative systems to making the specific decisions for which their accuracy has been empirically validated.

This example also illustrates dangers inherent in mistaking the plausibility of associations in interpretable systems for causal relationships that can be exploited through intervention. Machine learning systems can leverage associations in data sets to make highly accurate predictions and diagnoses. Except for systems specifically designed for causal discovery,[27] it is a mistake to expect those associations to track causal relationships in a way that we can exploit through intervention. Interpretability may thus feed a misguided expectation that understanding a set of associations valuable for specific diagnostic or prediction tasks will increase our ability to perform additional tasks to which those associations are not well suited and for which their accuracy has not been validated.

We saw earlier that one reason explanation is seen as the hallmark of expertise is that it involves communicating causal relationships in the relevant domain to stakeholders. When we lack causal knowledge in a domain, however, systems that use complex associations to reliably make diagnostic or predictive decisions with high sensitivity and specificity can have significant value. Because those associations do not directly track causal relationships, the value of interpretability is not clear.

It is also unclear what interpretability amounts to. Human decisions are often interpretable in the sense that we can rationalize them after the fact. But such rationalizations don't necessarily reveal why a person made the decision, since the same decision may be open to many different post-hoc rationalizations. As Zachary Lipton has argued,[28] machine learning systems are often interpretable in this sense as well.[29] If

this is all that is required to satisfy the requirement of interpretability, then both humans and machine learning systems are interpretable. We would therefore lack grounds for preferring less accurate models. But interpretability of this kind is unlikely to facilitate the goal of maintaining system reliability.

Interpretability might mean, instead, that humans should be capable of simulating the model a system uses for decision-making. This might involve taking "input data together with the parameters of the model and in reasonable time step[ping] through every calculation required to produce a prediction."[30] In this case, complex machine learning techniques, such as those used by deep learning systems, are currently uninterpretable. But so are some otherwise simple analytical approaches (such as linear models or rule-based systems) in sufficiently complex cases.[31] Most human decisions are not interpretable in this sense either. Given the degree of incompleteness in our own domain knowledge and the fact that associations do not necessarily capture causal relationships, it is not clear that the ability to "step through" a model will provide marginal improvements in reliability sufficient to offset marginal losses in accuracy. It may, however, lead overconfident analysts to use these models for purposes to which they are inherently unsuited.

Accountability and Nondomination

In productive sciences where domain models robustly capture causal relationships, the accuracy and reliability of diagnostic or prognostic decisions can be grounded in explanations that rely heavily on relationships and reasons derived from those models. In spheres where this knowledge is incomplete and piecemeal, the warrant for causal claims and assurances of accuracy and reliability must be grounded in empirical testing.

To promote accountability and to ensure that machine learning systems are not covert tools for arbitrary interference with stakeholder autonomy in medicine, regulatory practices should establish procedures that limit the use of machine learning systems to specific tasks for which their accuracy and reliability have been empirically validated. This promotes accountability and freedom from domination by enabling stakeholders to intelligently use artificial intelligence systems to perform tasks for which they are the most efficacious alternative—even if the grounds for their superior performance remain opaque.

To create such a system, greater emphasis should be placed on ensuring that data sets and analytical approaches are aligned with the decisions and uses they are intended to facilitate. Much as we seek to clarify the indications for which a drug can be prescribed, the use cases to which a machine learning system is suited and for which its accuracy and reliability have been validated should be clearly designated. Likely uses for which system performance has not been validated should be discouraged. The robustness of systems should be tested during development by exploring windows of operation outside of which accuracy and reliability break down. This involves validating system performance on multiple data sets that reflect the diversity of real-world contexts. Before deployment in clinical practice, system performance should also be tested against standard-of-care alternatives in prospective trials measuring impacts on clinically meaningful end points.[32] This means, in part, that when machine learning systems perform the same decision task as humans, the relative accuracy and reliability of humans and machines should be evaluated in well-designed empirical studies. Deployment should also include a plan for continuous quality improvement in which system performance can be audited and accuracy reassessed in light of changing clinical contexts.

Recommendations to prioritize explainability or interpretability over predictive and diagnostic accuracy are unwarranted in domains where our knowledge of underlying causal systems is lacking. Such recommendations can result in harms to patients whose diseases go undiagnosed or who are exposed to unnecessary additional testing. They may also encourage the use of machine learning systems for purposes to which they are not suited if associations in highly predictive models are mistakenly treated as causal relations that can be exploited through intervention without first validating the causal relevance of such associations.

Acknowledgments

I thank Jonathan Kimmelman for sage advice about relevant examples, David Danks for critical feedback on several drafts of this paper, and an anonymous referee for helpful suggestions.

1. N. Rotstein, "Challenges in Machine Learning: Cracking the Black Box Open," Medium, April 19, 2018, https:medium.com/maria-01/ok-computer-but-why-dde64c22b7ba.

2. W. R. Swartout, "XPLAIN: A System for Creating and Explaining Expert Consulting Programs," *Artificial Intelligence* 21, no. 3 (1983): 285-325.

3. S. Athey, "Beyond Prediction: Using Big Data for Policy Problems," *Science* 355, no. 6324 (2017): 483-85; see also D. E. Adkins, "Machine Learning and Electronic Health Records: A Paradigm Shift," *American Journal of Psychiatry* 174, no. 2 (2017): 93-94; R. Caruana et al., "Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York: Association for Computing Machinery, 2015), 1721-30.

4. Swartout, "XPLAIN," 286-87.

5. Ibid., 287.

6. Aristotle, *The Complete Works of Aristotle*, ed. J. Barnes (Princeton, NJ: Princeton University Press, 1984), *Metaphysics*, Ll.98lal2-24, and *Nicomachean Ethics*, VI.vii.

7. A. J. London, "Moral Knowledge and the Acquisition of Virtue in Aristotle's *Nicomachean* and *Eudemian Ethics*," *Review of Metaphysics* 54, no. 3 (2001): 553-83.

8. D.-H. Ruben, *Explaining Explanation* (New York: Routledge, 2015), 109.

9. W. C. Salmon, *Scientific Explanation and the Causal Structure of the World* (Princeton, NJ: Princeton University Press, 1984).

10. C. C. Fu and S. Wang, *Computational Analysis and Design of Bridge Structures* (Boca Raton, FL: CRC Press, 2014).

11. P. Petit, "Civic Republicanism," (Oxford: Oxford University Press, 1997).

12. V. Gulshan et al., "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus

Photographs," *Journal of the American Medical Association* 316, no. 22 (2016): 2402-10.
13. A. Rajkomar et al., "Scalable and Accurate Deep Learning with Electronic Health Records," *Digital Medicine* 1, no. 1 (2018): article 18.
14. Z. C. Lipton, "The Mythos of Model Interpretability," arXiv preprint, arXiv:1606.03490, 2016, https://arxiv.org/pdf/1606.03490.pdf.
15. M. Bayati et al., "Data-Driven Decisions for Reducing Readmissions for Heart Failure: General Methodology and Case Study," *PloS One* 9, no. 10 (2014): e109264.
16. J. Kimmelman and A. J. London, "Predicting Harms and Benefits in Translational Trials: Ethics, Evidence, and Uncertainty," *PLoS Medicine* 8, no. 3 (2011): e1001010.
17. J. Kimmelman and A. J. London, "The Structure of Clinical Translation: Efficiency, Information, and Ethics," *Hastings Center Report* 45, no. 2 (2015): 27-39.
18. D. W. Thomas et al., *Clinical Development Success Rates 2006–2015* (San Diego, CA: Biomedtracker, 2016).
19. "Carotene and Retinol Efficacy Trial (CARET)," National Cancer Institute, accessed December 20, 2018, https:epi.grants.cancer.gov/Consortia/members/caret.html.
20. Aristotle, *Nicomachean Ethics,* I.vii.
21. Ibid., VI.vii.
22. A. J. London and J. Kimmelman, "Accelerated Drug Approval and Health Inequality," *JAMA Internal Medicine* 176, no. 7 (2016): 883-84; see also Kimmelman and London, "Structure of Clinical Translation."
23. A. J. London, "Groundhog Day for Medical Artificial Intelligence," *Hastings Center Report* 48, no. 3 (2018): inside back cover.
24. Caruana et al., "Intelligible Models."
25. F. Cabitza, D. Ciucci, and R. Rasoini, "A Giant with Feet of Clay: On the Validity of the Data That Feed Machine Learning in Medicine," in *Organizing for the Digital World* (New York: Springer, 2019), 121-36; see also D. Danks and A. J. London, "Algorithmic Bias in Autonomous Systems," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (Marina del Rey, CA: IJCAI, 2017), 4691-97.
26. Caruana et al., "Intelligible Models."
27. P. Spirtes et al., *Causation, Prediction, and Search* (Cambridge, MA: MIT Press, 2000).
28. Lipton, "The Mythos of Model Interpretability."
29. Rajkomar et al., "Scalable and Accurate Deep Learning."
30. Lipton, "The Mythos of Model Interpretability."
31. Ibid.
32. M. D. Abràmoff et al., "Pivotal Trial of an Autonomous AI-Based Diagnostic System for Detection of Diabetic Retinopathy in Primary Care Offices," *NPJ Digital Medicine* 1, no. 1 (2018): article 39.

The Strange Tale of *Three Identical Strangers*: *Cinematic Lessons in Bioethics*

BY BRYANNA MOORE, JEREMY R. GARRETT, LESLIE ANN McNOLTY, AND MARIA CRISTINA MURANO

Tim Wardle's 2018 documentary film *Three Identical Strangers* is an exploration of identity, family, and loss. It's also about nature versus nurture and the boundaries of ethically permissible research, particularly research involving children. The film tells the story of identical triplets—David Kellman, Bobby Shafran, and Eddy Galland—who were separated soon after birth in 1961. A different family adopted each boy, without being told that their son had two identical brothers. Through sheer coincidence, at age nineteen, Bobby and Eddy met. After a local newspaper picked up their story and published a picture of them, David entered the fray. Their unlikely reunion became a national feel-good sensation. Then the real story began to unfold.

Bryanna Moore, Jeremy R. Garrett, Leslie Ann McNolty, and Maria Cristina Murano, "The Strange Tale of *Three Identical Strangers*: Cinematic Lessons in Bioethics," *Hastings Center Report* 49, no. 1 (2019): 21-23. DOI: 10.1002/hast.974

The adoption agency responsible for finding the families was collaborating with a group of researchers working on a study about . . . something. The design, purpose, and findings of the study, headed by Austrian psychiatrist and psychoanalyst Peter B. Neubauer, remain unpublished and are not exactly clear. In addition to sharing some of the personal trials and tribulations of the brothers and their families, the film details their fight to obtain information about the study and for closure. But while the film may have received rave reviews, *Three Identical Strangers* left us feeling uneasy. We worried that, with so much information about the study missing, any analysis of the film might fall prey to some of the same kinds of concerns that we had with the film. *Three Identical Strangers* demands ethical scrutiny, however.

We came away with two sets of questions. One set—and in a way, perhaps the simpler set—has to do with the story that the film documents. Separating biologically related children for the purpose of medical research may strike many as ethically problematic regardless of whether the researchers