by: Natalie Ram

Natalie Ram[1]

Modern algorithmic design far exceeds the limits of human cognition in many ways. Armed with large data sets, programmers promise that their algorithms can better predict which prisoners are most likely to recidivate[2] and where future crimes are likely to occur.[3] Software designers further hope to use large data sets to uncover relationships between genes and disease that would take human researchers much longer to identify.[4]

But modern machine learning still cannot effectively match human cognition in at least one crucial respect: learning from small data sets. Young children, for example, master new concepts with startling rapidity and fluency. "[G]iven 2 or 3 images of an animal you have never seen before, you can usually recognize it reliably later on."[5] Similarly, "a person only needs to see one Segway to acquire the concept and be able to discriminate future Segways from other vehicles like scooters and unicycles."[6] And as any parent of a toddler can attest, "children can acquire a new word from one encounter."[7]

Psychologists believe that, by the time we reach six years of age, "we recognize more than 10^4 categories of objects."[8] By contrast, traditional algorithmic design typically requires many more training examples. "[L]earning one object category requires a batch process involving thousands or tens of thousands of training examples."[9] Researchers describe the human method of learning new categories, and objects within categories, from one or a handful of examples "one shot learning."[10]

by: Natalie Ram

In a relatively recent body of work, researchers are beginning to take aim at cracking this insight of human learning—and teaching algorithmic systems to learn in the same way. This work is still developing,[11] with early examples demonstrating that one shot learning algorithms can correctly categorize, for example, human faces, motorbikes, airplanes, and spotted cats on par with big data algorithms while using only fifteen training examples.[12] While differentiating between such wildly disparate categories may seem a long ways away from facial recognition or other big data applications, research into such uses is already underway. Using a variety of algorithmic approaches, researchers have already demonstrated that one shot learning may enhance algorithmic applications in fields including facial recognition,[13] hand writing identification,[14] and shoe tread analysis.[15]

Yet, to date, legal academics have overlooked these efforts. Indeed, the phrase "one shot learning" does not appear in any law review article searchable in Westlaw.[16] This article seeks to remedy that gap in the literature, introducing the concept and language of one shot learning, and warning that enabling computer systems to successfully perform one shot learning is likely to exacerbate problems of insufficient transparency and of hidden bias that already beset the use of algorithmic systems, particularly in the criminal justice context.

Introducing One Shot Learning

One shot learning builds on, and learns from, traditional big data algorithmic design. In traditional big data algorithms, developers train a machine learning system to recognize and accurately distinguish a particular category by feeding the system thousands, and often tens of thousands, of examples of the relevant

by: Natalie Ram

category.[17] These training data often derive from records of past human conduct or from data coded by humans, which makes their development "a tedious and expensive task."[18] Moreover, new categories must be learned afresh, with many thousands of training examples for each new category.[19]

One shot learning, by contrast, attempts to replicate the human ability to apply past knowledge to new learning.[20] As one set of authors has explained, "[t]he key insight is that, rather than learning from scratch, one can take advantage of knowledge coming from previously learned categories, no matter how different these categories might be."[21] That is, once a machine learning system has learned a few categories "the hard way"—based on thousands of training examples—some "general knowledge" can be extracted and applied to new, previously unknown categories.[22] Thus, one shot learning algorithms are designed to make inferential leaps and to extract knowledge learned about one category to aid in identification of future categories.[23] In this way, one shot learning models aim to encode algorithmic systems with the power to learn how to learn.

Transparency and Understanding in One Shot Learning

The promise of one shot learning to enable algorithmic systems to learn new categories more cheaply and efficiently than in the past is enormous; but there is also significant risk that developments in one shot learning will exacerbate some of the most persistent difficulties in AI markets, including transparency.

Transparency—and the related problem of understanding—is a challenge for machine

by: Natalie Ram

learning models in at least two ways. First, the nature of much machine learning is opaque.[24]It may be formally opaque in instances where it is "actually impossible to state how the algorithm classifies observations once it has been developed."[25] In other instances, an algorithmic pattern will be functionally opaque, as where algorithmically identified relationships are "so complicated that they defy explicit understanding."[26] In both senses, machine-learning models act as a "black box,"[27]in which known data goes in, answers are produced, but the process by which data is transformed into answers is unknown and potentially unknowable. Under such circumstances, effective transparency and understanding may be difficult to achieve.

Second, these difficulties of transparency and understanding are exacerbated by the outsized role that trade secrecy occupies for complex algorithmic systems. In recent years, the Supreme Court has reinforced that mathematical processes are patent-ineligible "abstract ideas," at least insofar as those ideas are not inventively applied in some real-world application.[28]Machine learning models, as fundamentally mathematical processes, typically are excluded from patent protection.[29] In the absence of such protection, trade secrecy has become a primary method for maintaining competitive advantage.[30]Unfortunately, trade secret protection, by definition, depends on continued secrecy; public disclosure destroys it.[31]

Reliance on trade secrecy, and opacity about how a particular machine-learning model reaches decisions, is likely to pervade one shot learning to an even greater extent than traditional big data algorithms. Because one shot learning extrapolates the skill of "learning how to learn" from underlying categories that may themselves teach unexplainable relationships, it builds opacity upon opacity.[32] Moreover, one shot learning is likely to multiply the sources of information about an algorithmic

by: Natalie Ram

system that must be known to replicate or understand its workings, making trade secrecy a more powerful tool for competitive advantage and a greater foil to transparency and understanding. Some algorithmic systems are explainable upon examination of source code.[33] But because machine learning models are often a "black box," examining source code may provide insufficient insight into their validity or reliability. Instead, adequate understanding of the algorithmic system may depend on having access to both source code and training data.[34]Accordingly, when algorithmic developers have invoked trade secrecy to withhold training data, as well as information about algorithmic design, the non-disclosure of either renders obscure the system as a whole.[35]

Creators of one shot learning models, in turn, may well be able to stymie transparency and understanding even if they disclose both source code and the limited training examples used for new learning categories. Again, because these models depend on prior learning in the traditional "big data" way, the absence of the more remote training data that armed a one shot learning model with its "general knowledge" may make understanding that model difficult, if not impossible. By proliferating the sources of information that may be withheld to maintain competitive advantage, one shot learning may pose a greater threat to transparency and understanding than even traditional "black box" big data models.

These failures of disclosure and transparency are likely to impose significant practical barriers to developing sufficiently accurate and reliable algorithms.[36]Secret code is often less good code, as secrecy may obstruct effective oversight of the reliability and validity of algorithmic tools.[37] This difficulty is particularly likely to arise in a burgeoning field like one shot learning algorithms, where different

by: Natalie Ram

researchers have adopted different approaches to solving the one shot learning problem.[38] In other contexts in which programmers use somewhat different mathematical models (or code for the same model differently), the consequence is that, in attempting to do the same thing, these competing tools sometimes yield different results from identical inputs.[39]

Moreover, "black box" algorithms are likely to be particularly problematic in some of the settings in which one shot learning algorithms are most desired, including law enforcement investigations. As described above, researchers are already working to develop effective one shot learning algorithms to perform facial recognition,[40] hand writing identification,[41] and shoe tread analysis.[42] While these types of analysis may be useful in multiple contexts, they are likely to be of particular interest to law enforcement. Indeed, one article exploring the implementation of one shot learning methods for shoe tread analysis focuses explicitly on the forensic use of such algorithms, explaining, "We investigate the problem of automatically determining what type (brand/model/size) of shoe left an impression found at a crime scene."[43]

The use of undisclosed algorithmic models in the criminal justice setting, in turn, is frequently problematic. In addition to practical concerns about the impact of trade secrecy on algorithmic design, secrecy and opacity surrounding criminal justice algorithms can raise significant constitutional concerns.[44] I have argued elsewhere that secret criminal justice algorithms are "at least in tension with, if not in violation of, defendants' ability to vindicate their due process interests throughout the criminal justice process, as well as their confrontation rights at trial."[45] As one shot learning algorithms perform more complex inferential tasks, access to algorithm design information is likely to be even more critical in the criminal justice field to

ensure that the design accurately yields reliable results and functions as intended. Trade secrecy threatens to undermine those goals.

Exacerbating Bias in One Shot Learning

In addition to intensifying reliance on trade secrecy, one shot learning in algorithmic design may also multiply the ways in which algorithmic design encodes bias—and not in a good way. In traditional big data algorithms, because training data often derives from records of past human conduct or from data coded by humans, bias in this human conduct can give rise to biased outputs from the algorithm.[46] "[T]raining data is often gathered from people who manually inspect thousands of examples and tag each instance according to its category. The algorithm learns how to classify based on the definitions and criteria humans used to produce the training data, potentially introducing human bias into the classifier."[47]

In the context of one shot learning, if categories learned the "hard way" are tainted with bias, this may similarly infect new categories learned by inference. Indeed, any such bias may be amplified where there are fewer, rather than more, training examples for a new category. Where only a few examples of a new category are available, more inferential leaps in learning are required, and so bias in human selection or coding of examples is likely to be aggravated.

More troubling, in attempting to create more "human" learning, programmers designing one shot learning algorithms may replicate crucial faults in human learning and decision making. Human learning achieves rapid categorization and decision making in part through reliance on cognitive short cuts. These short cuts—called heuristics—operate as "principles by which [human beings] reduce the complex tasks

by: Natalie Ram

of assessing likelihoods and predicting values to simpler judgmental operations."[48] Heuristics, in other words, help human beings make inferential leaps from incomplete data. As Tversky and Kahneman have noted, "[i]n general, these heuristics are quite useful, but sometimes they lead to severe and systematic errors."

Among the most significant heuristics of human decision making is the representativeness heuristic. Humans intuitively call on this heuristic when tasked with assessing: "What is the probability that an object A belongs to a class B? What is the probability that event A originates from process B? What is the probability that process A will generate an event B?" Under the representativeness heuristic, these "probabilities are evaluated by the degree to which A is representative of B, i.e.. by the degree of similarity between them." The more similar A is to B, the higher the probability that A belongs to (or originates from) (or will generate) B. Typically, representativeness is a useful heuristic, and its probabilities are usually accurate enough for everyday living.[49]

But representativeness can also lead decision makers astray. For instance, in assessing whether Mr. X holds a particular occupation, decision makers assess "the similarity of Mr. X to the stereotype of each occupational role, and orders the occupations by the degree to which Mr. X is representative of these stereotypes." (Similarly, in assessing the likelihood that Mr. X will be a repeat criminal offender, decision makers assess how alike Mr. X is to the stereotype of a repeat criminal offender.) Yet stereotypes are, by definition, inexact. Moreover, these determinations of probability are frequently immune to crucial factors like, for instance, the base-rate of each occupation in the general population. Moreover, individuals express confidence in their predictions in accordance with the degree of

similarity between Mr. X and their stereotype of a particular profession, "with little or no regard for the factors that limit predictive accuracy."[50]

One shot learning algorithms appear to attempt to instruct machine systems to make the same sorts of inferential leaps and extrapolation from prior information that give rise to heuristics like representativeness in humans. These algorithms seek to make computer systems more human-like in their capacities for data analysis and recognition. But an unintended consequence of accomplishing this result may be to inflict on computer systems the same kinds of heuristics that render human decision making irrational in systematic ways. If successful, one shot learning may prove to be faster than human judgment, but perhaps not better than it.

Conclusion: Mapping the Solution Space

One shot learning advances machine learning by enabling sophisticated models to learn new categories from only a few examples. So long as the model has gained prior exposure to extensive training data about some categories, a one shot learning model can learn how to learn. That is exciting, as it opens the door to making greater use of more diverse data for machine learning.

But one shot learning also threatens to worsen already persistent problems in machine learning, including the opacity of machine learning models, their reliance on trade secrecy, and the bias they may unwittingly encode. These problems are not straightforward to solve. Moreover, these problems are likely to fester together, in that non-transparent systems shielded from effective external validation are less likely to recognize and correct for their unintended biases.

But trade secrecy, and the understanding and fairness it may threaten, need not

dominate innovation in one shot learning algorithms. Alternative mechanisms for innovation policy abound, including prizes, grants, regulatory exclusivities, and tax incentives.[51] These tools of innovation policy can help to support the research, development, and sale of effective one shot learning algorithms in place of (or alongside) trade secrecy. In particular, at this early stage in the development of one shot learning algorithms, grants and research-based tax incentives may be well suited to driving the disclosure of early-stage work related to both algorithmic design and training data—and thus to driving a swifter pace of innovation in the field more broadly. Grants and tax incentives effectively infuse investment dollars in research up front, rather than rewarding the successful completion of a commercial product.[52] In so doing, these innovation policy levers "may enable more and smaller companies to enter the market," as they reduce "the private capital investments required for innovation."[53]

Moreover, policymakers may also drive development of valid and reliable one shot learning models by investing in complementary incentives for innovation. In discussing innovation incentives for developing valid and reliable black box algorithms in the health care context, Nicholson Price has suggested that "direct or indirect government intervention could usefully aid the generation of datasets," to be used as common infrastructure for algorithmic development.[54] Such a solution to problems of data secrecy and fragmentation is particularly well suited to one shot learning in the criminal justice context, as government will often have a monopoly on the data necessary for training these algorithms at the outset.[55] After all, government actors are responsible for the investigation, arrest, prosecution, and incarceration of criminal defendants in the United States, and so have unique access to data about these populations.[56] Similarly, government is well positioned to incentivize innovative

methods for validating black box algorithms through prizes or grants for outside validation of complex algorithmic models, including those involving one shot learning.[57]

Complex algorithmic systems, including those deploying one shot learning, hold enormous promise for expanding the range of data from which an algorithmic system can learn and the range of categories it can learn to identify. But this promise is not unfettered. If these complex models are to be deployed, particularly in the criminal justice context, relevant stakeholders—including policymakers, courts, prosecutors, and defense counsel alike—must grasp the ways in which machine learning in general, and one shot learning in particular, may undermine as well as enhance the pursuit of justice and take steps to mitigate those harms.

By Natalie Ram, Assistant Professor, University of Baltimore School of Law

1. Assistant Professor, University of Baltimore School of Law; J.D., Yale Law School; A.B., Princeton University. Many thanks to the participants of UCLA Law School's conference on AI in Strategic Context: Development Paths, Impacts, and Governance, and in particular to Richard Re for excellent comments on earlier drafts of this piece.
2. *See, e.g.*, Julia Angwin et al., *Machine Bias*, ProPublica (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (describing COMPAS, one risk assessment software package).
3. Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 Stan. L. Rev. 1343, 1368 (2018) ("Predictive policing systems . . . may rely on historical data to model the likelihood of future

crimes . . . .").

4. *See, e.g.*, W. Nicholson Price II, *Black-Box Medicine*, 28 Harv. J.L. & Tech. 419 (2015) (describing the promise and difficulties of highly complex predictive algorithms for personalized medicine).
5. Li Fei-Fei, Rob Fergus & Pietro Perona, *A Bayesian Approach to Unsupervised One-Shot Learning of Object Categories*, Proc. of the Ninth IEEE Int'l Conf. on Computer Vision (2003).
6. Brenden M. Lake et al., *One Shot Learning of Simple Visual Concepts*, 33 Proc. of the Annual Meeting of the Cognitive Sci. Soc'y 2568, 2568 (2011).
7. *Id.*
8. Li Fei-Fei, Rob Fergus & Pietro Perona, *One-Shot Learning of Object Categories*, 28 IEEE Transactions on Pattern Analysis and Machine Intelligence 594, 594 (2006).
9. *Id.*
10. *See, e.g.*, *id.*; Lake et al., *supra* note 5.
11. *See* Gregory Koch, Richard Zemel & Ruslan Salakhutdinov, Siamese Neural Networks for One-Shot Image Recognition, Proc. of the 32nd Int'l Conf. on Machine Learning (2015) ("Overall, research into one-shot learning algorithms is fairly immature and has received limited attention by the machine learning community.").
12. Fei-Fei et al., *supra* note 7, at 594.
13. *Id.*
14. *See, e.g.*, Brenden M. Lake et al., One Shot Learning of Simple Visual Concepts, 33 Proc. of the Annual Meeting of the Cognitive Sci. Soc'y 2568 (2011).
15. Bailey Kong et al., Cross-Domain Forensic Shoeprint Matching, British Machine Vision Conference (2017).
16. Searching "one shot learning" and "'one shot' w/3 learning" in "All Content" database on Westlaw Next, with zero relevant results. http://next.westlaw.com (last visited Oct. 25, 2018).

17. *See* Fei-Fei et al., *supra* note 4, at 1.
18. *Id.*
19. *Id.*
20. *See* Fei-Fei et al., *supra* note 7, at 594.
21. *Id.*
22. *Id.*
23. *See, e.g.*, Fei-Fei et al., *supra* note 7, at 594 (implementing an algorithmic approach designed to "make use of the knowledge that has been gained so far rather than starting from scratch each time we learn a new category").
24. *See* W. Nicholson Price II, *Big Data, Patents, and the Future of Medicine*, 37 Cardozo L. Rev. 1401, 1407 (2016).
25. *Id.* at 1410.
26. Price, *supra* note 23, at 1410.
27. *See* Price, *supra* note 3, at 441 (describing these black box phenomena in the context of medical innovation).
28. Alice Corp. Pty. v. CLS Bank Int'l, 134 S. Ct. 2347, 2357 (2014); *see also* Ass'n for Molecular Pathology v. Myriad Genetics, Inc., 569 U.S. 576, 589 (2013); Mayo Collab. Servs. v. Prometheus Labs., Inc., 566 U.S. 66 (2012); Bilski v. Kappos, 561 U.S. 593 (2010).
29. *See* Natalie Ram, *Innovating Criminal Justice*, 112 Northwestern U. L. Rev. 659, 703 (2018).
30. *See* Wexler, *supra* note 2, at 16-17; *see also* Price, *supra* note 23, at 1407.
31. W. Nicholson Price II, *Regulating Secrecy*, 91 Wash. L. Rev. 1769, 1776-77 (2016).
32. *See* Fei-Fei et al., *supra* note 7, at 594.
33. *See* State v. Chun, 943 A.2d 114, 159 (N.J. 2008) (identifying source code errors in the Draeger Alcotest alcohol breath test device following independent source code examination, but nonetheless concluding that the Alcotest was reliable); Charles

Short, Note, *Guilt by Machine: The Problem of Source Code Discovery in Florida DUI Prosecutions*, 61 Fla. L. Rev. 177, 185 (2009) (discussing *Chun* and the source code errors that independent analysis uncovered in the Draeger Alcotest).

34. *See* Laurel Eckhouse et al., *Layers of Bias: A Unified Approach for Understanding Problems with Risk Assessment*, Crim. Justice & Behavior, Nov. 2018, at 17 ("Transparency in both risk scoring and training data is a necessity for researchers to be able to vet risk-assessment instruments.").
35. *See* Wexler, *supra* note 2, at 1374 n. 162 ("Transparency in both risk scoring and training data is a necessity for researchers to be able to vet risk assessment instruments." (internal quotation marks omitted)).
36. Ram, *supra* note 28, 686-90 (describing the importance of access to code to ensure the accuracy and reliability of algorithmic design, and the tension between trade secrecy and code quality).
37. *Id.*
38. *Compare* Fei-Fei et al., *supra* note 7, at 594 (adopting a "Bayesian" approach) and Fei-Fei et al., *supra* note 4 (utilizing a "Bayesian framework"), *with* Koch et al., *supra* note 10, at 3 (utilizing a "siamese convolutional neural network"). *See generally* Koch et al., *supra*, at 2-3 (describing lines of research in one shot learning that have adopted various approaches).
39. *See* Ram, *supra* note 28, at 681-82.
40. *See* Fei-Fei et al., *supra* note 7, at 594.
41. *See* Lake et al., *supra* note 5.
42. *See* Kong et al., *supra* note 14.
43. *Id.* at 1.
44. *See* Ram, *supra* note 28, at 692-99.
45. *Id.* at 692.
46. *See* Wexler, *supra* note 2, at 1348.

47. Nicholas Diakopoulos, Algorithmic Accountability: On the Investigation of Black Boxes 6 (2014), https://towcenter.org/research/algorithmic-accountability-on-the-investigation-of-black-boxes-2/.
48. *See* Amos Tversky & Daniel Kahneman, *Judgment under Uncertainty: Heuristics and Biases*, 185 Science 1124 (1974).
49. *Id.*
50. *Id.*
51. *See, e.g.*, Ram, *supra* note 28, at 704-714 (describing alternative mechanisms to trade secrecy for incentivizing algorithmic innovation); Daniel J. Hemel & Lisa Larrimore Ouellette, *Beyond the Patents-Prizes Debate*, 92 Tex. L. Rev. 303 (2013) (summarizing the literature on patents versus prizes versus grants and adding tax incentives to the range of innovation policy levers).
52. *See* Ram, *supra* note 28, at 707-09 (discussing grants), 712-13 (discussing tax incentives).
53. *Id.* at 709.
54. Price, *supra* note 23, at 1440.
55. *See, e.g.*, Northpointe, Practitioners Guide to COMPAS 2 (2012) ("The updated normative data were sampled from over 30,000 COMPAS assessments conducted between January 2004 and November 2005 at prison, parole, jail and probation sites across the United States.").
56. Ram, *supra* note 28, at 709.
57. *See* Price, *supra* note 23, at 1451 (describing "validation bounties" in the context of "black box medicine"); Ram, *supra* note 28, at 718 (describing the need for broad disclosure of algorithmic information to "enable[] multiple groups, including nonprofit criminal defense organizations, to share the financial and other costs of validating a software program and examining software updates and software status

on an ongoing basis").