The Stakes of Uncertainty: Developing and Integrating Machine Learning in Clinical Care

MADELEINE CLARE ELISH
Data & Society Research Institute

The wide-spread deployment of machine learning tools within healthcare is on the horizon. However, the hype around "AI" tends to divert attention toward the spectacular, and away from the more mundane and ground-level aspects of new technologies that shape technological adoption and integration. This paper examines the development of a machine learning-driven sepsis risk detection tool in a hospital Emergency Department in order to interrogate the contingent and deeply contextual ways in which AI technologies are likely be adopted in healthcare. In particular, the paper bring into focus the epistemological implications of introducing a machine learning-driven tool into a clinical setting by analyzing shifting categories of trust, evidence, and authority. The paper further explores the conditions of certainty in the disciplinary contexts of data science and ethnography, and offers a potential reframing of the work of doing data science and machine learning as "computational ethnography" in order to surface potential pathways for developing effective, human-centered AI.

INTRODUCTION

"The problem with sepsis is that we just don't know," says an ER physician who is explaining the process of treating sepsis in her Emergency Department, "Everything is risk versus benefit… we're hoping that Sepsis Watch will help us get to the right point faster." Sepsis Watch is a machine learning-driven system that assesses a patient's risk of developing sepsis, an extremely deadly syndrome that involves a body's over-response to infection and likely results in organ damage or death if untreated. The system, or "the tool" as the clinicians and computer scientists describe it, is one of the first machine learning models for detecting sepsis to be deployed in an Emergency Department (ED).[1] After years of development, Sepsis Watch is in the process of being integrated in the context of emergency care at Duke University Hospital, a large, urban research hospital.

In this paper, I bring an anthropological perspective to bear on the development of Sepsis Watch. My focus is on articulating the epistemological and social entanglements that characterize the emergence of this technology, and analyzing the implications of these entanglements for conceptions of trust, evidence, and authority in clinical care. I begin by providing an overview of the project background, and then move into a discussion of the technologies at stake. I then describe and discuss a set of tensions that emerged from research during the development and planning stages of the tool. In the final section, I draw together the themes of trust, evidence, and authority with the disciplinary groundings of machine learning and ethnography, and propose the idea that machine learning may be productively understood as "computational ethnography." My aim is to explore a set of tensions that arise around certainty, explainability, process, and method, and in turn to offer a potential reframing of the work of doing data science and machine learning in order to surface potential pathways for developing effective, human-centered AI.

This research is based on an ongoing collaboration with a multi-disciplinary team at Duke Health and Duke Institute for Health Innovation (DIHI) working together to design, develop, and implement the tool. The observations described in this paper are based on approximately 20 hours of interviews with technologists, clinicians, and administrators, and approximately 15 hours of observation in the Emergency Department and Cardiac Intensive Care Unit at Duke. In writing this paper I also draw on over three years of participant observation of machine learning technologies in a range of sectors as a researcher at the Data & Society Research Institute.

BACKGROUND

The problem: Sepsis

Sepsis is a widespread and grave problem in healthcare. Some research concludes that over 3.1 million hospitalizations annually are due to severe sepsis, and the U.S. Center for Disease Control (CDC) reports that 1.5 million people develop sepsis annually, and about 250,000 people die from sepsis within the United States each year (CDC 2018). The CDC reports that one out of every three patients who die in a hospital have sepsis.

Sepsis is a syndrome characterized by a body's extreme response to an infection, and without treatment leads to tissue damage, organ failure, and death. Although the condition is more likely to develop in populations with reduced immune responses, sepsis may affect anyone. Not only is the condition life-threatening, it also develops rapidly, often in a matter of hours. Early detection and treatment are critical for survival (Levy et al. 2010). However, there does not exist one test to confirm the onset of sepsis.

Clinicians attempt to diagnose sepsis by detecting an underlying infection that can be tested through blood samples in a lab. However, these tests are not always reliable and may take too long; for instance, blood culture results are updated at 24 and 48 hour intervals. Although the timely diagnosis of sepsis is extremely challenging, once it has been diagnosed, treatment is usually straight-forward and based on existing protocols of antibiotics, known as bundles, that have proven effective in treating sepsis.[2]

Sepsis is a particularly relevant disease to discuss in the context of categories of evidence and certainty. Sepsis is extremely difficult to diagnose in large part because its causes and progression are very poorly understood from a biomedical perspective. As one emergency doctor explained, "Sepsis is very difficult. … how do you know when someone has it? 'I just know' isn't good enough as a diagnostic tool — it's not like an x-ray, where we can see, 'oh that's broken.'" Diagnosing sepsis, in practice, is often ultimately left to "gut instinct" as another clinician told me.

Taken together, these characteristics make sepsis an ideal case for machine learning diagnostics, "prime for tech" as one clinician told me. Sepsis is widespread, with profound severity for people's lives—and also for the allocation of hospital and insurance resources. One emergency doctor explained how "sepsis is always on the front burner." First and foremost, he explained, there's "the moral and ethical imperative" to save people's lives, and at the same time, "there's the brick and mortar bottom line." A technological tool that could facilitate diagnosis is very needed and would be very welcome across the board.

The technology: Machine learning and AI

Most of my informants and collaborators use the term “machine learning” or “deep learning” when referring to Sepsis Watch and their field of research or work. Several doctors I spoke with emphasized that it was best to avoid terms like AI or machine learning when talking to doctors, favoring terms like “predictive analytics.” Nonetheless, these terms exist within a constellation of technologies implicated in the umbrella term of AI, a term that is both over-hyped and unmistakably compelling. As the latest technology buzzword to enter common parlance, part of AI’s rhetorical power is in its slipperiness, wherein everyone has a notion of what AI is—but everyone’s notion is different (Elish and Hwang 2015).

Even as AI is a nebulous term—more marketing than technical—machine learning does refer to a specific set of computer science and statistical techniques that refer to a type of computer program or algorithm that enables a computer to “learn” from a provided dataset and make appropriate predictions based on that data. Computer scientist Tom Mitchell’s definition makes clear what “learning” means in this context: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E" (Mitchell 1997: 6). Learning, in this case, is narrowly defined and refers essentially to the capacity for an algorithm to recognize a defined characteristic in a dataset in relation to a defined goal and to improve the capacity to recognize this characteristic by repeated exposure to the dataset.

For the purposes of this paper, it is relevant to call attention to precisely what is “intelligent” in machine learning and current formulations of AI. Until relatively recently, the intelligence at stake in “AI” predominantly referred to procedural, logic-based reasoning and the capacity to manipulate abstract symbolic representations – now also known as “Good, old fashioned AI” (Hoagland 1995). These systems formed the commercial technologies within the “expert systems” of 1970s and 1980s, which were eventually critiqued as “brittle” and too limited (Dreyfus 1972; Forsythe 1993, 2002; Suchman 2007). This stands in contrast to the “intelligence” at stake in current conceptions of AI, which are rooted in machine learning techniques, in which logic and abstract representational meaning are beside the point; intelligence is derived through detecting patterns across vast amounts of data and predicting outcomes based on probabilistic statistics. In other words, the “smartness” of AI comes from a system’s ability to process and analyze huge amounts of data, beyond the scale of any individual human, in order to predict or automate certain activities. The datasets and models used in these systems are not objective representations of reality. A machine learning algorithm can be said to “know” something only in the sense that it can correlate certain relevant variables accurately. These different paradigms of intelligence within AI research have deep implications for the construction of knowledge, truth, and fact as epistemic categories and the ways in which these categories can be leveraged in social practice.

DEVELOPMENT AND DEPLOYMENT

The team: Duke Health and DIHI

The interdisciplinary team working on Sepsis Watch is led by both physicians and computer scientists and is split across Duke Medicine and the Duke Institute for Health Innovation

(DIHI). DIHI is a multidisciplinary institute that draws together both Duke University and Duke Medicine. According to their website, DIHI "promotes innovation in health and health care through high-impact innovation pilots, leadership development, and cultivation of a community of entrepreneurship" (DIHI 2018). With a staff of approximately fifteen, DIHI functions as a kind of collaboratory, leading or facilitating pilot projects across Duke, involving clinicians and students at Duke Health as well as faculty, staff, and students at Duke University. While each of the existing twenty projects is unique, all are "grassroots projects" in the sense that they were first proposed by external collaborators and then developed in partnership with the team at DIHI.

In the previous section, I described the problem that Sepsis Watch has been developed to address, and then outlined the relevant technological context. This order reflects the ways in which the Sepsis Watch team working on this project approaches their work. It is notable that this is the opposite of dominant norms around developing new AI interventions. In the current climate of AI hype, it is common for new companies or projects to be technology-driven, as opposed to problem- or community-driven. As ethnographers, we know the perils of this approach. The DIHI team is also wary of tech-solutionism, and prioritizing clinical problems and institutional goals is structured into their project conception and development process. For instance, potential projects need to start from a real problem with a specific goal – not an idea for a model that can predict something and see if it will work in real life.

The tool: Sepsis Watch

The origins of Sepsis Watch date back several years to a Duke hospital initiative aimed at improving patient outcomes and decreasing costs of care. However, the project in this iteration began in 2016 when two physicians wrote a small proposal to work with DIHI on the project. Once initiated, the first twelve months of the project involved obtaining and cleaning data from a database of Electronic Health Records (EHRs).

The road to deployment was rockier and slower than anticipated. While the team worked closely with The Epic Systems Corporation, simply referred to as Epic, the leading provider of EHRs to US hospitals, to extract data and explore potential development paths, the model was not able to be deployed within Epic software for the duration of the pilot. Another set of unexpected problems and delays emerged around integration of the tool within existing IT infrastructures. In this context, it is interesting to contrast the Silicon Valley ideal of disruption with the reality of large, legacy healthcare systems, which I discuss in further detail below.

The tool itself leverages real-time data drawn from EHRs, a deep learning model, and a graphical user interface (GUI) in order to predict the risk of a patient developing sepsis before it occurs. The training dataset for the model consisted of 51,697 inpatient admissions at Duke Hospital spanning the course of 18 months stored in Duke's Epic HER (Futoma et al. 2017: 6). This patient data was considered to be representative of the hospital's patient population and clinical settings, and 80% of the data was used for training the model, with the remaining 20% held out for multiple stages of verification. The output of the model produces a score and the score determines whether a patient is "no risk," "low risk," "high risk," or "septic" based on streaming data including lab results, vitals, and medications (Futoma et al. 2017). The GUI allows a clinician to keep track of several patient scores at a time, and when desired, pull up a specific patient's score over time, as well as preceding treatment, and what and when the last lab, vital, and medication were taken. The clinician, after reviewing the scores, can send the patient to "a watch list," and monitor the watch list, which keeps track of all the treatments (completed or uncompleted), chart reviews, and clinical encounters. The application will be accessed through a tablet or other handheld device, a design requirement from the beginning of the project that derived from the need for nurses to remain mobile and able to move around the hospital

Acknowledging limits: "Where can we have a real effect?"

Healthcare, and in particular hospitals, have historically been slow to adopt new technologies was emphasized to me again and again during interviews, and is often a talking point in healthcare industry conferences and op-eds. This is a point of frustration for many, who see the use of big data and machine learning in healthcare as providing immense opportunity to improve patient outcomes. Both industry and academic researchers are focused on developing new models or tools. Using big data in healthcare, for instance in problems like identifying sepsis may seem like a low-hanging fruit, but as one informant put it: "It's a low hanging fruit, but the fruit has a thick stem. You can't really hit it."

An important part of the development process for the team has been learning to work in different timeframes, not just for developing a data pipeline or building a model, but also testing and implementing the tool in the emergency department; most previous projects had been one year long and discreet pilot projects, but this project was intended to operate a different scale of complexity and deployment. The team, perhaps because they are situated within a research institute and teaching hospital, is aware and open to the potential that the tool may not work or be accepted by clinicians. The incentives for the team are aligned not around profit or technical success, per se, but rather, around patient outcomes and basic research. In order to give the tool the best chance of formal success, the team carefully articulated what they thought was a feasible goal for the project: to manage the care of sepsis better – not necessarily to decrease mortality. In addition, when planning for the tool's integration into existing workflows and adoption, the team was careful to explicitly bound the capabilities and role of the tool. In reading a previous version of this paper, one of the leads on the project commented that

> Our experience testing the model during the silent period…has reinforced this. All that we know now is that the model works at predicting the first episode of sepsis in patients admitted to Duke University Hospital, but before transfer to an Intensive Care Unit. Going to a different hospital or expanding into different inpatient settings will all require additional work to validate that the tool continues to perform as well as we expect it.

The choice of the word "tool" to describe the technology was unintentional, but nevertheless underscores the idea of *augmenting* existing clinical practice – not replacing.

SOCIOTECHNICAL ENACTMENTS

In this section, I discuss three salient dimensions of developing Sepsis Watch as they emerged from interviews and ethnographic research and that bring into focus the epistemological implications of introducing a machine learning-driven risk assessment tool into a clinical setting: trust, evidence, and authority. While on the surface these categories may seem relatively self-evident, it is precisely in the articulation of just what these concepts mean that the cultural work and social infrastructures of deploying machine learning systems emerge. I follow Anne-Marie Mol's articulation of "enactments" in clinical settings in order to draw attention to "the techniques that make things visible, audible, tangible, or knowable" (2002: 33). Things, in this case, can be diseases or bodies or any manner of object, and enactments allow us to talk about the ways that realities are multiple but rooted in praxis. That realities are enacted – not *constructed* or *performed* in the senses in which these terms have become overloaded in STS (Science and Technology Studies) literature -- "enacted" draws attention to the particular reconfigurations of the world that are not given or even self-evident but rather emerge as significant at particular times and places among specific actors. In this context, I discuss the ways in which trust, evidence, and authority are talked about and enacted alongside and through the machine learning tool. Far from simple categories, the analysis demonstrates that the tool, as a machine learning technology, requires careful negotiations and interpersonal and institutional reconfigurations even before it can be fully deployed.

Trust: "The answer to trust is not a technical solution"

The nature and development of "trust" was a recurring theme in my conversations with all members of the team. All the team leaders knew that establishing "trust" was an essential foundation upon which everything else would rest. Only if a technology is trusted will it be used. They judged that the successful deployment of Sepsis Watch would rely not only within the clinical interactions in the emergency department, but also in the various formal and informal social networks that intersect, shape, and use or are responsible for the use of the tool. These networks would set the stage for whether people would trust and accept the technology. While the lead physicians were attuned to the importance this "change management," the machine learning researchers perhaps underestimated the extent to which actually deploying the tool would all too often feel like a Sisyphean task.

How does one build a technology that can be trusted? A primary strategy of the team was to "loop in stakeholders" from the very beginning of developing the tool. This included conversations and meetings with not only with hospital leadership, but also physicians, nurses, and other front-line workers who would be using the tool, as well as all the departments that would eventually need to be involved, including the hospital IT department—"the guys who keep the lights on." For each of these stakeholders, it was important to tread carefully and not be or *be seen as* telling other people how to do their work. At the same time, emphasizing the ways in which the tool fit into existing standardized guidelines for care and also could potentially improve patient outcomes was a generative form of engagement.

This strategy was of course informed by previous work and established best practices. As one physician put it:

> When you're in charge of a life, I think this level of distrust is an important one to have until it's validated. … Any time you are adopting new technology which is not validated, I think there is some amount of trust building that has to go along with the project and that comes from working with an engagement right from the beginning.

But in the lead up to the testing, no one knew how to ensure that trust would be established and this caused anxiety.

Another significant concern revolved around the risk of "alarm fatigue." A previously implemented early warning system (Bedoya at al. 2018) to identify patients at risk of cardiac arrest, unplanned ICU admission, and death based on the National Early Warning Score (NEWS) (Smith et al 2013) within the Duke University Health System resulted in 63.4% of alerts triggered to be dismissed by the care nurse who was notified (Futoma et al. 2017). This previous system was a much less sophisticated and precise system than Sepsis Watch. Still, concerns about "alarm fatigue" were a common theme in my interviews with clinicians. Alerts, like pop-up windows on a personal computer, are often experienced as more annoying than helpful, like the "update software and restart" alerts that seems to show up constantly in the corner of a screen on Windows and Mac operating systems. The tricky but necessary balance to achieve, a doctor explained, speaking about alert systems generally, is that "at some point, someone has to write a rule to reach a threshold [to trigger the alert]. If it's too low, it'll get ignored cause it'll alert too often. And if it's too high, you're risking harm to a patient." Establishing trust around a tool is not only about building inter-personal relationships but also about aligning desired behaviors to existing norms and expectations.

Evidence: "Our machine learning is easy to call a black box—but the human body is a black box!"

Key to having meaningful interactions with stakeholders was demonstrating evidence that the tool would be effective. What could constitute evidence or even efficacy was simultaneously central and variable. What kind of evidence matters? Where does it come from and who can interpret it? An important insight of the team leading up to testing was that various types of evidence are salient for different stakeholders: Hospital administrators and managers are convinced by numbers and statistical trends. Front-line clinicians and middle managers are more convinced through anecdotal evidence and discussions of specific cases and patient outcomes. This is not just because it's a compelling story, but because through relating the specifics clinicians have the possibility to identify with a mistake or an oversight they themselves might have made along the way. Telling these stories and going over cases are built into recurring department meetings and are part of student training.

In addition to the *types* of evidence that were salient, the team also had discussions about the *extent* of evidence necessary. Doctors, for instance, people explained to me, are trained to look beneath the surface and understand cause and effect; how much will doctors need or want to look inside the black box of a machine learning tool?

While it is tempting to speak generally about one category of ER physician, for instance, the distinct nature of different types of hospitals makes this problematic. During a tour I was given of the Emergency Department, my guide told me laughing, "If you've seen one academic hospital, you've seen one academic hospital." In the community hospitals that are part of the Duke network but are less prestigious, sometimes with more limited resources, and also with different communities for whom they provide care, physicians and nurses were thought to have different requirements. Existing patterns of knowledge diffusion mean that "the Duke brand" plays a large role in how much people may trust something coming from the research hospital. Duke's reputation and history profoundly influence how people will perceive a technology coming out of a Duke research center.

Still, the working understanding was that many physicians will only trust a machine learning model if they have proof that it works. The researchers felt there was an important distinction between *proving* that a model works and proving *how* it works. They felt, in fact, that because the model was uninterpretable they had to be "even more rigorous" in how the model was tested. The team published technical papers and spent significant time trying to demonstrate that their model performed "better than the status quo" within and beyond the hospital. The technical lead on the project stated, "If our model didn't perform better than every comparable model on our held out sets, there's no amount of trust we could have tried to build via relationships." Both demonstrations of efficacy and enforcing social networks were deemed necessary to establish evidence, a classic theme in the history of science and technology (Shapin and Schaffer 1985).

Interestingly, this troubles a growing emphasis on explainability and interpretability in technical and social science research communities.[3] What does it mean to look into the black box, if everyone has different conceptual lenses through which to see what's inside? Making a model "technically interpretable," the machine learning researchers emphasized, does not equate to make the technology interpretable or trusted by doctors. As one researcher put it: "I think the issue is that interpretability is about understanding the *causation*. That's the key thing that people push for, but instead they would say, 'I want to be able to interpret the model.'" The Sepsis Watch model is "totally uninterpretable" but their development process focused on the trust that can be built from technical demonstrations of efficacy embedded within existing social relationships.

Sepsis is a particularly interesting syndrome in which to think about tradeoffs around interpretability. As described earlier, the causes of sepsis are poorly understood. Other ethnographic work by Maiers (2017) about automated decision aids for detecting sepsis in infants describes the diagnosis of sepsis as being built on "a gut feeling." Moreover, treatment is not "path dependent;" that is, once a patient has sepsis the treatment is not dependent on how sepsis developed. However, this is not the case for all conditions. For example, treatment for cardiogenic shock, a condition in which the heart cannot pump enough blood, is dependent upon what caused the shock. During one conversation a researcher exclaimed, "Our machine learning is easy to call a black box--but the human body is a black box!" Sepsis is like a black box, inside another black box.

Other physicians voiced the opinion that if it seemed to be working by an agreed upon metric, in the case of Sepsis Watch, improving care for septic patients, the inner workings of the model didn't matter. Following one conversation, a researcher directed me to one of his favorite TED talks by Ziad Obermeyer (2017) on the subject, answered the question, "If a machine could predict your death, should it?" with a resounding yes. In this talk, Obermeyer emphasized that the implications for living life well and dying a good death—and the limitations and failures of existing healthcare—were too profound to be discounted when judging the risks and benefits of using a technology without understanding how precisely it works.

Authority: Who's responsible? Who's accountable? Who needs to be communicated with, and who needs to be informed?

As voiced above, the distrust that clinicians have of new technologies are well founded and reasonable. The life of a patient is in a clinicians' hands. Trust has been placed in them and their judgement, and it is their responsibility and professional duty to ensure the patient receives the best care possible.

Lines of authority became relevant on multiple levels during the development process. On one level, the team needed to understand who held ultimate authority to allow the implementation of the tool in the first place. In the context of a vast and slow-moving industry, with intricate policies around data privacy and security, the sources of authority were multiple and not always easy to see.

On another level, the everyday practices of clinicians might be variously enhanced, threatened, or destabilized in the face of a new machine learning tool. One physician emphasized the fear that many doctors have that machine learning and AI will threaten their "autonomy." She explained,

> A lot of this predictive models have the perception of taking away some decisional timing, like you're doing algorithmic or robot-based medicine. Machine learning itself, the term, you're not supposed to use it as much when talking to physicians because it's got a negative connotation.

She said the preferred term right now when talking to clinicians is "predictive analytics." Physicians were also worried about how the jobs of nurses would change. One concern was that nurses would need to learn to interpret and work with the tool in ways that they had not been trained for. Previous studies have shown mixed results (Guidi 2015) when implementing diagnostic aides or automated decision tools in clinical settings. Maiers (2017) argues that nurses in a Neonatal Intensive Care Unit (NICU) incorporate a predictive analytics tool into sepsis diagnosis and care through interpretive processes that combine multiple forms of evidence, including experiential and embodied knowledge. And while in some cases, nursing staff have experience the health information technology tools as empowering and enhancing their abilities to do their work (de Vries et al. 2017), other studies have found the emergence of "data silos" (Leslie et al. 2017), which we might understand as coinciding with what Fiore-Gartland and Neff (2015) termed as "data valences," calling attention to the distinct interpretations and expectations about the same data that different actors may have in healthcare settings.[4] Supporting this, when asked about the potential utility of machine learning tools to assist in diagnosis, a nurse stated concisely, "Your numbers are only as good as the one who's interpreting them." This nurse also felt that such tools would probably be more useful to newer nurses, and less useful for those with more patient care experience.

While it is essential that we explore how new modes of sense-making are emerging, and creating new epistemic paradigms, we must also examine the new over-reaches and blind spots that accompany such shifts. In this final section, I discuss the themes above and propose that the disciplinary evolution of anthropology and ethnography might be leveraged to reframe data science and the ways in which new "intelligent" technologies are deployed.

REFRAMING THE WORK OF BUILDING AI

Machine Learning as Alchemy

One of the largest gatherings of machine learning researchers is a conference called NIPS, the conference on Neural Information Processing Systems. During the NIPS 2017 conference, which was attended by over 7,000 people (Gershgorn 2017), a long-standing and well-respected member of the NIPS community and current research scientist at Google named Ali Rahimi gave one of the most talked about presentations. During his acceptance speech for an award given for a lasting contribution to the field , he provocatively argued that "machine learning has become alchemy" (Rahimi 2017). This indictment of the field relied on mobilizing the wide-spread understanding of alchemy as a "pseudo-science" – an ancient art built on occult knowledge and superstition. Although the history of alchemy is more complex (Moran 2006), alchemy is commonly perceived as the antithesis of the modern principles of science and reproducible experimentation.

To call a roomful of computer scientists with advanced degrees "pseudo-scientists" was quite a blow. The proposed parallel was that many of today's machine learning models, especially those that involve the use of neural nets or deep learning, are poorly understood and under-theorized; the outputs are correct even if the mechanisms work are unknown.[5] Advances tend to occur more through trial and error than theoretical developments.[6] Rahimi emphasized that while sometimes it may not matter much, when such systems are in charge of people's lives and livelihoods in domains like healthcare and criminal justice, this may be unacceptable. Rahimi concluded: "I would like to live in a world whose systems are built on rigorous, reliable, verifiable knowledge, and not on alchemy."

While Rahimi's talk was widely discussed and well-received, it was not without its detractors. In a Facebook post responding to Rahimi, Yann LeCun, Chief AI Scientist at Facebook, wrote,

> It's insulting, yes. But never mind that: It's wrong! … Sticking to a set of methods just because you can do theory about it, while ignoring a set of methods that empirically work better just because you don't (yet) understand them theoretically is akin to looking for your lost car keys under the street light knowing you lost them someplace else. Yes, we need better understanding of our methods. But the correct attitude is to attempt to fix the situation, not to insult a whole community for not having succeeded in fixing it yet. (LeCun 2017)

The two sides represented by Rahimi and LeCun draw out a fundamental tension in science and engineering: must theory come before practice? Is one more valuable than the other? These are questions with deep epistemological implications: how do we know what we know? What claims to truth are we able to make? Why does it matter?

In the context of healthcare, these questions also have life or death implications. Reflecting on these issues, a Sepsis Watch researcher observed,

> There is nothing in medicine comparable to Newton's Laws of Motion that has stood the test of time for approximately 350 years. What has stood the test of time is the Hippocratic Oath, which concerns ethics, not knowledge. The state of knowledge is constantly evolving and even knowledge that seems reliable and verifiable at one point is rapidly debunked.

As the TED talk by Ziad Obermeyer (2017) referenced early asked, "If a machine could predict your death, should it?" If it meant you could live a better life, do you need to know why?

Just as calls for algorithmic transparency gave way to calls for explainable and interpretable machine learning after critical communities realized that transparency itself is not an end goal, but a means to a goal, so it seems explainable and interpretable machine learning must also be thought of a means, not a goal in and of itself. Members of the Sepsis Watch team voiced the opinion that when people talk about "interpretability" what they really are talking about is causality. The Sepsis Watch machine learning model is not interpretable, and while they have developed the GUI to display particular readings that the model indicates are out of the ordinary, causality is not indicated. Recall: the precise causes of sepsis are still unknown. As discussed above, the team considered the role of trust as key to addressing the motivations behind the interest in interpretability; the tool was developed to be in close alignment with federal guidelines around sepsis treatment, and also over years in collaboration with prominent Duke physicians who will also be overseeing its testing and deployment.

But proponents of the need for explainable and auditable AI and machine learning raise important considerations. The implications of black box algorithms of all kinds for legal due process and accountability (Citron 2008; Crawford and Schultz 2013; Pasquale 2015) are troubling and leave open the door for intentional as well as unintentional unfair discrimination and unequal opportunities to resources and care. Checks on the expectations of machine learning systems to assess, recommend, or decide may be grappled with on specific technical teams, as has been the case with Sepsis Watch, but broader understandings of these limits need to be developed and established more widely.

Machine Learning as Computational Ethnography

Data science and the knowledge it produces are often asked to play the role of objective quantifier (Beer 2016), presenting the cold, hard facts. This perception of immutable truth is a surface to productively crack. Elsewhere, danah boyd and I have proposed that one way to ground the universalizing claims of data science would be to develop a rich methodological reflexivity like that at the heart of ethnography, embracing the partiality and situatedness of data science practice (Elish and boyd 2017). We proposed that machine learning could be seen as a form of computational ethnography. The comparison of anthropology and data science is not as odd as it might seem; Like ethnographers, data scientists immerse themselves with data ("a field site"), selecting data points or patterns from what Bronislaw Malinowski, a founding figure of ethnographic methods, once termed, "the imponderabilia of actual life" (Malinowski 1984: 18). They select from a plethora of data a smaller subset that they find significant based on their intuition and training, and then iteratively develop models and frameworks to fit or explain their findings. Over decades, anthropology as a discipline has developed a foundation of methodological reflexivity, confronting the limits of its own knowledge production, ranging from the articulation of research agendas and areas of focus (Asad 1973; Faubion & Marcus 2009; Hymes 1974) to the cultural and geographical delineations of those areas (Gupta & Ferguson 1997), to the very modes of representation and engagement at stake in ethnographic research (Cefkin 2010; Clifford & Marcus 1986; Taussig 2011). The invocation of ethnography is a means to open up the possibilities of contextualizing what it means to produce knowledge about the world and developing a discipline that can grapple with an iterative and interpretive way of knowing.

Expecting Uncertainty

Contextualizing the insights of machine learning systems as situated and partial is key not only to developing the field of computer science *but also* to facilitating effective integration into everyday work practices. This is especially true in the healthcare context. In a review of studies examining unintended consequences of machine learning in medicine published in the *Journal of the American Medical Association*, Cabizta (2017) argues that machine learning-based decision support systems may be problematic because they "bind empirical data to categorial interpretation" (E2), and require "considering digital data as reliable and complete representations of the phenomena" (E1). In the race to optimize and personalize, the variability of interpretations and the diverse contexts of health data and health care may get lost (Ferryman and Pitcan 2018). He draws on one study to demonstrate the ways in which a loss of context produced a technically valid data model but one that incorrectly predicted mortality risk (Caruana et al. 2015); the model predicted (and it seemed counter-intuitive but was judged to be accurate at the time) that patients with both pneumonia and asthma were at a lower risk of death from pneumonia than patients with only pneumonia. Ultimately, the researchers realized this was the case because patients who had a history of asthma and came to the hospital with pneumonia were usually admitted directly to an intensive care unit, which led to better outcomes. When the model was built, this institutional context and behavior was not represented in the model. Cabizta concludes:

> Users and designers of ML-DSS [machine learning decision support systems] need to be aware of the inevitable intrinsic uncertainties that are deeply embedded in medical science. Further research should be aimed at developing and validating machine learning algorithms that can adapt to input data reflecting the nature of medical information, rather than at imposing an idea of data accuracy and completeness that do not fit patient records and medical registries, for which data quality is far from optimal. (E2)

Like sepsis, the extent of the problem with a data model might not reveal itself plainly at the beginning. This warrants the need for acknowledging limits, iterative development, and vigilant testing throughout a model's deployment. Embracing the contingencies and incompleteness of a machine learning model is a way not only to make a better tool, but also to fit its development within its future use contexts.

CONCLUSION

From Startup to Endups

This paper has highlighted a set of interwoven stories that are implicated in the integration of a machine learning risk-detection tool. All are related to the shifting grounds of evidence and certainty in the context of machine learning systems. One storyline has been about the development of a machine learning technology in the clinical context of an Emergency Department at Duke University Hospital. The development and eventual integration of Sepsis Watch in the ED has brought bring into focus the epistemological implications of introducing a machine learning-driven tool into a clinical setting by analyzing shifting categories of trust, evidence, and authority. These shifts speak to the complex set of social practices that are necessary to ensure effective clincial care even as a machine learning system introduces an automated "intelligent" actor into the clincial setting (Lustig et al. 2016).

Another storyline has about the growth and development of disciplinary evidence. This story, more abstract and historical in nature, brings to bear the history of ethnography on the development of data science. While machine learning and ethnography are often thought of as at odds, they share many orientations toward collecting data and inductively piecing together coherent wholes from minute particulars. What would it mean for machine learning, and data science more generally, to adopt a reflexive and situated epistemic posture? How might this posture better take into account the social contexts within which machine learning technologies are deployed, and the ways in which they might be most effectively integrated into existing work practices?

Taken together, these stories offer insight into the changing nature of evidence, how it is constituted, and how it is recognized as meaningful. Through an analysis of how trust, evidence, authority, and the limitations of technologies are "enacted" (Mol 2002) during the development of a machine learning technology, this paper hopes to have contributed to understanding how machine learning will alter the field of health care and medical expertise, and also how machine learning, as a mode of knowing, will alter how we make sense of the world.

While so many resources and so much energy are focused on creating, finding, or investing in the right start-ups, we need more people to be thinking about "end-ups" (Maeda 2013). How do technologies move from prototype to institutional process? This is the real and hard work of building AI and machine learning technologies. As the case of Sepsis Watch demonstrates, the path is far from straight-forward. Focusing only on a technical functionality is not enough. In addition to technical research that focuses on robust functionality and clinical research that focuses on patient outcomes, *socio-technical* ethnographic research is necessary to understand and plan for the ways in which technologies will be disruptive or effectively integrated into society.

M.C. Elish is a Research Lead and co-founder of the AI on the Ground Initiative at Data & Society Research Institute, an independent non-profit research institute focused on the social implications of data-centric technological development. She received her PhD in Anthropology from Columbia University, and an M.S. in Comparative Media Studies from MIT. mcelish@datasociety.net

NOTES

Acknowledgments – This research is indebted to Mark Sendak and the DIHI Sepsis Watch team for insightful conversations and sharing their time with me. I am grateful to Ian Lowrie, Patrick Davison, and Kadija Ferryman for their feedback on previous drafts. This research was conducted with generous financial support from the AI Ethics and Governance Fund. All oversights, errors, and omissions are entirely my own.

[1] During a recent machine learning for healthcare conference, a tweet from the conference organizers put it plainly: "We cannot underemphasize that this is a concrete end-to-end deployment of an accurate deep learning clinical predicion [sic] model for real-time patient monitoring. NO ONE ELSE IS DOING THIS YET! #MLHC2018 #ml4healthcare."

[2] Bundles in healthcare refer to a set of predetermined clinical elements of care based on evidence-based practice guidelines and which have been demonstrated to work effectively in concert. The treatment of sepsis with bundles, and the adoption of related protocols as formal federal and state regulation was largely driven by the global health initiative, "Surviving Sepsis Campaign" (2002) which began in 2002.

3. For instance, major technical conferences like NIPS and ICML now have devoted symposia and sessions for interpretable AI, and the explainable AI has been a growing focus of inter- disciplinary conferences FAT* (Fairness, Accountability, and Transparency) conference and the law and technology conference, WeRobot. The Defense Research Projects Agency (DARPA) has opened a program for Explainable AI research.

4. See also previous older work from EPIC 2012 about the implications of implementing EMRs for divisions of labor within clinical settings (Vinkhuyzen, Plurkowski, and David 2012).

5. For an overview of the controversy and differing viewpoints at stake, see Peng (2017).

6. This is not to say that machine learning and data science researchers do not draw on existing theories or attempt to develop what they consider rigorous knowledge practices (Lowrie 2017).

REFERENCES CITED

Asad, Talal
1973 Anthropology & the Colonial Encounter. Ithaca, NY: Ithaca Press.

Beane, Matthew
2018 Shadow Learning: Building Robotic Surgical Skill When Approved Means Fail. *Administrative Science Quarterly* 33(January), 000183921775169–37.

Bedoya, Armando, Meredith Clement, M. Phelan R.C. Steorts, Cara O'Brien, and B.A. Goldstein
2018 Minimal Impact of Implemented Early Warning Score and Best Practice Alert for Patient Deterioration. *American Journal of Respitory and Critical Care Medicine* 197:A4295.

Beer, David
2017 The Social Power of Algorithms. *Information, Communication & Society*, 1–13.

Burrel, Jenna
2016 How the Machine "Thinks:" Understanding Opacity in Machine Learning Algorithms. *Big Data & Society* 3(1), 1-12.

Cabitza, Federico, Raffaele Rasoini, and Gian Franco Gensini
2017 Unintended Consequences of Machine Learning in Medicine. *JAMA* 318 (6), 517–2.

Caruana R, Lou Y, Gehrke J, et al.
2015 Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721-1730.

CDC
2018 Sepsis > Data & Reports Center for Disesase Control and Prevention. https://www.cdc.gov/sepsis/datareports/index.html (Aug 24, 2018).

Cefkin, Melissa
2010 *Ethnography and the Corporate Encounter: Reflections on Research in and of Corporations*. New York: Berghahn.

Clifford, James, and George E. Marcus
1986 *Writing Culture: The Poetics and Politics of Ethnography*. Berkeley, CA: University of California Press.

Citron, Danielle
2008 Technological Due Process. *Washington University Law Review* 85, 1249-1313.

Crawford, Kate and Jason Schultz
2013 Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. *Boston College Law Review* 55, 93-128.

de Vries, Aisha, Jos M T Draaisma, and Joris Fuijkschot.
2017 Clinician Perceptions of an Early Warning System on Patient Safety. *Hospital Pediatrics*, September, 0138–10.

DIHI
2018 Home page. Duke Institute for Health Innovation. https://dihi.org/ (Aug 24, 2018).

Dreyfus, Herbert
1972 *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, MA: MIT Press.

Dummett, B Alex, Carmen Adams, Elizabeth Scuth, Vincent Liu, Margaret Guo, and Gabriel J Escobar.
2017 Incorporating an Early Detection System Into Routine Clinical Practice in Two Community Hospitals," December, 1–7.

Elish, M.C. and danah boyd
2017 Situating Methods in the Magic of Big Data and AI. *Communications Monographs* 85:1, 57-80.

Elish, M.C. and Tim Hwang
2016 *An AI Pattern Language*. New York: Data & Society Research Institute.

Ferryman, Kadija and Mikaela Pitcan
2018 *Fairness in Precision Medicine*. New York: Data & Society Research Institute.

Fiore-Gartland, Brittany, and Gina Neff
2015 Communication, Mediation, and the Expectations of Data: Data Valences Across Health and Wellness Communities. *International Journal of Communication*, May, 1466–84.

Faubion, James and George E. Marcus
2009 *Fieldwork Is Not What It Used To Be: Learning Anthropology's Method in a time of Transition*. Ithaca, NY: Cornell University Press.

Forsythe, Diana
2002 *Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence*. Stanford, CA: Stanford University Press.

1993 Engineering Knowledge: The Construction of Knowledge in Artificial Intelligence. *Social Studies of Science* 23(3), 445-477.

Futoma, Joseph, Sanjay Hariharan, Katherine Heller, Mark Sendak, Nathan Brajer, Meredith Clement, Armando Bedoya, and Cara Obrien
2017 An Improved Multi-Output Gaussian Process RNN with Real-Time Validation for Early Sepsis Detection. *Proceedings of Machine Learning for Healthcare* 68 (August), 1–12.

Gershgorn, Dave
2017 The world's biggest and most important AI conference. Quartz.com, Dec 12. https://qz.com/1152156/inside-the-worlds-biggest-and-most-important-ai-conference/ (Aug 24, 2018).

Guidi, Jessica L, Katherine Clark, Mark T Upton, Hilary Faust, Craig A Umscheid, Meghan B Lane-Fall, Mark E Mikkelsen, et al.
2015 Clinician Perception of the Effectiveness of an Automated Early Warning and Response System for Sepsis in an Academic Medical Center. *Annals of the American Thoracic Society* 12 (10): 1514–19.

Gupta, Akhil & Ferguson, James
1997 *Anthropological Locations: Boundaries and Grounds of a Field Science.* Berkeley, CA: University of California Press.

Haugeland, John
1985 *Artificial Intelligence: The Very Idea.* Cambridge, MA: MIT Press.

Hymes, Dell
1974 *Reinventing Anthropology.* New York: Vintage.

LeCun, Yann
2017 My take on Ali Rahimi's "Test of Time" award talk at NIPS. Facebook.com/yann.lecun. Facebook post. https://www.facebook.com/yann.lecun/posts/10154938130592143 (August 24, 2018).

Leslie, Myles, Elise Paradis, Michael A Gropper, Simon Kitto, Scott Reeves, and Peter Pronovost
2017 An Ethnographic Study of Health Information Technology Use in Three Intensive Care Units." *Health Services Research* 52 (4), 1330–48.

Levy MM, Dellinger RP, Townsend SR, Linde-Zwirble WT, Marshall JC, Bion J, Schorr C, Artigas A, Ramsay G, Beale R, et al.
2010 The Surviving Sepsis Campaign: results of an international guideline-based performance improvement program targeting severe sepsis. *Intensive Care Medine*, 36, 222–231.

Lowrie, Ian
2017 Algorithmic rationality: Epistemology and efficiency in the data sciences. *Big Data & Society* (Jan-June), 1-13.

Lustig, Caitlin, Katie Pine, Bonnie Nardi, Lilly Irani, Min Kyung Lee, Dawn Nafus, and Christian Sandvig
2016 Algorithmic Authority. *CHI* , 1057–62.

Maeda, John
2013 Startups are great, but we can learn a lot from "end-ups," too. Gigaom.com, Feb 3. https://gigaom.com/2013/02/03/we-all-love-start-ups-and-sometimes-forget-we-can-learn-a-lot-from-end-ups-too/ (Aug 24, 2018).

Maiers, Claire
2017 Analytics in Action: Users and Predictive Data in the Neonatal Intensive Care Unit. *Information, Communication & Society* 20 (6), 915–29.

Malinowski, Branislaw
1984 *Argonauts of the Western Pacific.* Prospect Heights, IL: Waveland Press.

Mitchell, Tom
1997 *Machine Learning.* New York: McGraw Hill.

Mol, Annemarie
2002 *The Body Multiple: Ontology in Medical Practice.* Durham, NC: Duke University Press.

Moran, Bruce
2006 *Distilling Knowledge: Alchemy, Chemistry, and the Scientific Revolution.* Cambridge, MA: Harvard University Press.

Obermeyer, Ziad
2017 If a Machine Could Predict Your Death, Should it? TEDxBoston Oct 24. https://www.youtube.com/watch?v=jeGJax4SLP0 (Aug 24, 2018).

Pasquale, Frank
2015 *The Black Box Society: The Secret Algorithms that Control Money and Information.* Cambridge, MA: Harvard University Press.

Peng, Tony
2017 LeCun vs Rahimi: Has Machine Learning Become Alchemy? *Synced: AI Technology & Industry Review.* https://medium.com/@Synced/lecun-vs-rahimi-has-machine-learning-become-alchemy-21cb1557920d (Aug 24, 2018).

Rahimi, Ali. 2017. Ali Rahimi's talk at NIPS (NIPS 2017 Test-of-time award presentation). Online video, https://www.youtube.com/watch?v=Qi1Yry33TQE.

Rhee, Chanu, Raymund Dantes, Lauren Epstein, David J Murphy, Christopher W Seymour, Theodore J Iwashyna, Sameer S Kadri, et al.
2017 Incidence and Trends of Sepsis in US Hospitals Using Clinical vs Claims Data, 2009-2014." *JAMA* 318 (13), 1241–49.

Smith, G.B., D.R. Prytherch, P. Meredith, P.E. Schmidt, and P.I. Featherstone
2013 The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation* 84(4): 465-70.

Shapin, Steve and Simon Schaffer
1985 *Leviathan and the Air-Pump: Hobbes, Boyle and the Experimental Life.* Princeton, NJ: Princeton University Press.

Stevens, Hallam
2017 A Feeling for the Algorithm: Working Knowledge and Big Data in Biology. *Osiris* 32(1), 151–174.

Suchman, Lucy
2007 *Human-Machine Reconfigurations: Plans and Situated Actions, 2nd ed.* New York: Cambridge University Press.

Surviving Sepsis Campaign
2002 "History" http://survivingsepsis.org/About-SSC/Pages/History.aspx (Oct , 2018).

Taussig, Michael
2011 *I Swear I Saw This.* Chicago: University of Chicago Press.

Vinkhuyzen, Erik, Luke Plurkowski, and Gary David
2012 Implementing EMRs: Learnings From a Video Ethnography. *EPIC Ethnographic Praxis in Industry Conference*, February, 235–48.