Developing AI for Weather Prediction: Ethics of Design and Anxieties about Automation at the US Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography

Przemyslaw Matt Lukacz

Department of the History of Science, Harvard University, United States/plukacz@g.harvard.edu

Abstract

The question of how professional and lay communities develop trust in new technologies, and automation in particular, has been a matter of lively debate. As a charismatic technology, artificial intelligence (A.I.) has been a common topic of these debates. This paper presents a case study of how the discourses and principles of ethics of technology development—specifically, of A.I.— were mobilized to form trust among actors in the fields of computer science, risk communication, and weather forecasting. My analysis draws on sociology of expertise and the literature on ethics of A.I. to ask: how emerging networks of expertise use ethics to overcome mistrust in technology? And, what role does the institutionalization of those networks play in the process of trust formation? I situate this discussion on the NSF Institute for Research on Trustworthy A.I. The Institute is positioned as a mediating organization with the goal of increasing trust in this technology primarily in the weather forecasting community, but also among the public. I show that first, to better understand how scientific and professional fields react to increased automation it is crucial to unpack the historical backdrop of how the professional identity of those experts has been shaped by a relationship with computer-supported modeling. To this end, I situate the discussion in the long-standing tensions between computer modelling and tacit knowledge in weather forecasting. Second, I argue that the means of establishing trust in A.I. propagated by the actors in the paper, which pair norms of explainability to sensitivity to professional intuitions and domain-specific conventions, rely on a series of 'mutual orientations' (Edwards, 1996). I mobilize the concept of 'mutual orientations' to describe the work of tailoring the ethics of A.I. to the specific requirements of weather sciences, but also to the vision of a national strategy of investment in this technology.

Keywords: Ethics of AI, Sociology of Expertise, Technological and Scientific Movements, Social Construction of Technology, Trust.

Introduction

Scientific expertise and moral values are intricately interwoven in the process of knowledge production (Shapin, 2008). Both federal funding agencies and academic researchers need to attest to the relevance of their technological and intellectual products by addressing the question of social relevance and ethical standards. This is the case with applied computational researchers. The question of how conceptions of ethics are legitimated and institutionalized gains critical importance during the contemporary acceleration of research on a form of artificial intelligence (AI) called machine learning (ML), as ML has proven to create epistemological and normative disruption in sciences and other professional fields (Kitchin, 2014). Ethics of design of algorithms and automation systems has been a site of an ongoing debate in and out of the academy (e.g., Dubber et al., 2020; Metcalf and Moss, 2019; Mittelstadt et al., 2016).

This paper explores the processes through which researchers working on applying ML to socially relevant issues legitimize and institutionalize their work by tailoring research to the emerging standards of ethics of AI. This discussion is empirically grounded in a partnership between AI researchers and weather forecasters. Weather forecasting is a generative site for theorizing how AI becomes embedded in scientific domains due to the long-standing tensions between computer modeling, automation, and tacit knowledge in the field—a jurisdictional struggle that AI has a capacity to exacerbate.

In this light, there is a need to understand better the processes through which ML and other forms of data-driven science become legitimized and institutionalized within domains where ML has previously played a marginal role. Furthermore, the paper asks: how do organizations involved in AI development resolve a tension between adapting external standards of ethics versus developing their own situated standards? I offer an analysis of the United States National Science Foundation's (NSF) Trustworthy AI Institute, which was established in 2020. The Institute has begun to develop ML for environmental sciences and weather forecasting. The bedrock of the Institute's operations is a design of ML that various communities of practice and potential users can trust. In this sense, the Institute's leaders attempt to frame the Institute as a mediating organization with the goal of increasing trust in the use of AI among environmental scientists, weather forecasters, and the public, but also to forge a closer partnership between the data analytics industry, government, and academia. This reworking of disciplinary and professional boundaries is centrally concerned with enabling innovation (Rottner, 2019) in the ability to predict future environmental conditions. The analysis of the Institute's multi-sector and multidisciplinary model is of relevance to the contemporary political and environmental milieu in which trust in the accuracy of weather and climate predictions is of high stakes.

The article explores how the conceptions of ethics, trust, algorithmic explainability, and adherence to the laws of atmospheric physics intersect in the design practices and discourse at the Institute. I show how a team of experts in AI, earth sciences, and risk assessment who were behind the Institute's formation set in motion a vision of the future of weather forecasting. This vision strives to fit into the prevailing imaginary of AI development and mitigate the mistrust in the technology among weather forecasters. Today's mistrust in AI on the part of the weather prediction community is a product of the long history of the external influence of computer science and modeling, which for over seven decades now has been shaping the identity of weather forecasters as an independent profession. I integrate a historical discussion into the article to locate the Institute's endevours within an established in the historiography of weather prediction theme of anxieties about automation and modelling.

The core theoretical contribution of the paper is a framework that describes stages of top-down and bottom-up 'mutual orientations' (Edwards, 1996) between a group of researchers and a federal funding agency towards institutionalization of an 'alternative expertise network' (Eyal, 2013) through reliance on a vision of civic, ethical, and trustworthy science. To do this, I appropriate the concept of 'mutual orientation' from a historian of science, Paul Edwards (1996). The concept means to capture a process of simultaneous alignment of objectives between a funding agency and a fundee. One of the reasons I choose to locate this inquiry on the case of implementation of AI in weather forecasting is the fact that the fears of "technological unemployment" (Keynes, 1930) are a long-standing issue in the history of weather forecasting (Harper, 2012). As such enduring conflicts between automation and tacit knowledge (Polanyi, 2009) can easily be ignored, one of the intended takeaways of the paper is to raise awareness about the need to consider disciplinary histories when analyzing the contemporary uptake of AI.

Methodology

To construct the paper's argument, I have relied on primary sources, including publicly available documents about the Institute (i.e., public presentations, online reports, calls for proposals) and peer-reviewed work of the Insititute's members. I transcribed recordings of Institiue-wide meetings, which I gained with the permission of the institute's directorate, and analyzed the transcripts according to the principles of content analysis. Furthermore, I used discourse analysis to capture policy discourses of the US AI strategy through reading NSF's and National Research Council's publications on trustworthiness. My reconstruction of the history of anxieties about automation in weather prediction was based on the reading of secondary literature.

Methodologically, I drew on a tradition of the social construction of technology (SCOT) (Pinch and Bijker, 1984) and the sociology of technological and scientific movements (TSIMs) and alternative expertise networks (Frickel and Gross, 2005). By adapting perspectives of the SCOT school of Science and Technology Studies (STS), I was able to examine processes of interpretative flexibility of trustworthy AI frameworks and the dynamics between designers and users of technology. Taking notice of the latter clarifications of the SCOT approach (Bijker, 1997), I attempted to pay special attention to shared 'technological frames' between designers and users of technology (trustworthy AI framework was one such framing device). I supplemented the SCOT methodology with a form of historical-sociological reconstruction of an alternative expertise network to capture its dynamic unfolding and a 'mutual orientation' towards an institutional (NSF's) vision of technology development.

This paper responds to scholarship parsing the problem of how trust is established between different groups of scientists and between users and designers of technology. The problem of mistrust among scientists most often emerges when groups of experts compete over the 'jurisdiction' (Abbott, 2014[1988]) for a specific task. Sociologists of expertise (Eyal, 2013) have asked how the 'jurisdictional struggle' between science and nonscience produces different forms of legitimation and institutionalization (Epstein, 1995, 2008; Gieryn, 1983; Star and Griesemer, 1989). I build on the contributions to this literature which focus specifically on how the creation of new and alternative expertise networks influences specific disciplines (Collins et al., 2007) and the problem of the jurisdictional struggle between scientific experts. Furthermore, to understand how the members of the Institute use trust as a 'boundary object' (Gieryn, 1983) in the process of institutional mutual orientations, I draw inspiration from the scholarship in science studies on the effect of organizations as meso-level structures which triangulate between scientific domains, national governments, and the industry (Vaughan, 1999; Guston, 1999).

Situating trust in sociology of expertise and social studies of algorithms

Trust is a key component of scientific practice (Shapin, 1994; Porter, 2020 [1995]). And while Anthony Giddens (1990) argued that trust is a defining feature of modernity, we live in an era of increasing mistrust in science (Eyal, 2019; Oreskes, 2019; Nichols, 2017). In the context of science done at the Institute, the question of trust manifests across three overlapping axes: trust between designers and users of new technology, between weather forecasters and AI, and between two epistemic communities (Knorr-Cetina, 1999): weather prediction and computational science. Examining these diverse dimensions of trust calls for synthesizing a few separate strands of debates in the social studies of algorithms.

So, why mistrust AI? The most well-known problem with AI systems is their 'black-boxed' character (Christin, 2020; Pasquale, 2015). The actors depicted in the following pages and researchers in many other domains are attempting to rectify precisely the problem of black boxing by creating 'explainable AI' (Hoffman et al., 2018). Explainability is but one of the examples of emerging conceptions of ethics and trust in AI. In this context, it has become customary for organizations concerned with AI development to publicize their value statements. Many of these frameworks share a core set of principles, or what Greene et al. (2019) called the 'moral background' (Abend, 2014) of AI value statements, which they define as "the grounding assumptions and terms of debate that make conversations around ethics and AI/ML possible in the first place" (Greene et al., 2019: 2122). The question behind this paper is: how does the moral background of AI development shape attempts at articulating situated, use-inspired, and domain-specific value frameworks?

Morality and trust in this context are two independent variables that feed into the same problem: ethics of design. Many authors in critical algorithm and data studies (see Illiadis and Russo, 2016; Moats and Seaver, 2019) have attempted to pin down what ethics both does and should imply (e.g., Richterich, 2018). Some authors have even unpacked the "ethics of ethics of AI" (Hagendorff, 2020; Powers and Ganascia, 2020). Drawing on the discussion about the ethics of algorithms by Mittelstadt and colleagues (2016), I understand the ethics of AI to imply two semi-distinct sets of concerns: epistemic and normative concerns. While the authors observe that "[d]istinct epistemic and normative concerns are often treated as a cluster" (Mittelstadt et al., 2016: 14), I concur that this strategy is analytically disadvantageous because the normative concerns often relate to the public perception and effects of technology, while the epistemic concerns are prioritized by the technology's users and designers. The analytical distinction helps to describe the details of the effects of ethical frameworks.

Mittelstadt et al.'s (2016) typology of standards of AI ethics lists 1) inconclusive evidence, 2) inscrutable evidence, 3) misguided evidence, as epistemic concerns. And 4) unfair outcomes, 5) transformative effects, as normative concerns, with traceability (the ability to determine wherein the process of design an "ethical bug" is embedded) as a technical concerns. In sum, Mittlestadt et al.'s typology gives precise language for inquiries into AI ethics. Nonetheless, I agree with the authors in stating that a "mature 'ethics of algorithms' does not yet exist, in part because 'algorithm' as a concept describes a prohibitively broad range of software and information systems" (Mittelstadt et al., 2016). As this case study shows, the development of ethical standards of AI needs to be grounded in and tailored towards the specific needs of professional communities.

The relationship between trust and ethics begs for more explanation. While both in this paper and in the literature on AI ethics at large, trust and ethics often appear together without much reflection, the actors depicted in the following pages adhere to the view that trust is a key component of an ethical AI. Interestingly, in part, because AI is most often anthropomorphized, the statement that AI should be trusted or trustworthy can be misleading. For example, Mark Ryan argues that "Overall, proponents of AI ethics should abandon the 'trustworthy AI' paradigm (…) replacing it with the reliable AI approach, instead," and adds that it should be the institutions using AI that should be trusted, and not the technology itself (Ryan, 2020: 17). Rather than resolving this definitional tension, my goal is to depict how the actors at the Institute follow similar to proposed by Ryan strategy by constructing a trustworthy and ethical institution in a form of collaborative research methodology closely aligned with the notion of 'AI ethics by design' (d'Aquin et al,, 2018).

Furthermore, the question of trust in AI relates to the problem of automation-led unemployment—one of the key themes in the social studies of algorithms (Benanav, 2020; Besteman and Gusterson, 2019; Eubanks, 2018; Ford, 2015). Much of this discussion centers on the perception of AI as a new and relatively obscure technology that engenders mistrust based on fear among many professionals about being replaced. After all, one could "trust" a technology –in the sense that it will work reliably—and also fear it. In fact, the more reliable the technology, the more one might fear that it will replace people. Resolving this tension is a non-trivial task. For example, Peter McClure (2018) links this 'technophobia' to a general lack of comprehension of the new technologies in a sizable portion of the U.S. population. McClure concludes that technological apprehension is exacerbated by fears of "technological unemployment" (Keynes, 1930; Floridi, 2014). The institute's focus on trust aims to, at the same time, mitigate fears of forecasters about being unemployed and allow them means of engaging with the design process of AI to make technology trustworthy through explainability and adherence to the laws of atmospheric physics.

Mutual orientations

Paul Edwards (1996) introduced the concept of 'mutual orientation' in chapter three of *The Closed World*. Edwards described how an early computer pioneer from MIT, Jay Forrester, convinced the U.S. military of the utility of digital computation to acquire funding for developing a general-purpose computer called "Whirlwind." Forrester's project had to compete with twelve other general-purpose digital computers funded by the Department of Defense. Therefore, Forrester had to present it as more urgent and critical than other early computers. Forrester and his group saw a potential application of their computer to the real-time military control system. Crucially, the focus on real-time control enabled by Whirlwind was the orientation Forrester chose to satisfy the granting agency's needs and compete with the larger pool of digital computer developers. In the words of Paul Edwards, "Forrester's (and MIT's) increasingly grand attempts to imagine military applications for Whirlwind represented expert 'grantsmanship,' or deliberate tailoring of grant proposals to the aims of funding agencies" (Edwards, 1996: 81). But Forrester also informed the funding agency about "as yet undreamt-of possibilities for automated, centralized command and control" (Edwards, 1996: 82). In effect, Forrester framed his research to suit the discourse of command and control, while the military embraced this imaginary as it was partially produced through Forrester's deliberate actions. Forrester's plan was met with strong resistance from the US generals, who saw the idea of being replaced by a computer as unacceptable. It was hence crucial for Forrester to promote trust in his automated technology.

The following analysis shows both mutual orientation between the NSF and the Institute (based on making AI trustworthy) and AI and weather prediction experts (based on making AI explainable and in alignment with the professional intuition of the forecasters). This article captures the following stages in this process: First, a scientific field (in this case, the field of ML) responds to social demands for a new ethical standard for innovation. Second, a network of applied computational researchers seeks collaboration with domain experts, leading to an alternative expertise network. Two kinds of mutual orientations then take place, and furthermore:

a. Mutual orientations between a scientific movement and a funding agency, enabled by a shared ethics of technology development, lead to institutionalization of the movement,
b. The institutionalization leads to the emergence and legitimation of new hybrid forms of expertise.

This theoretical framing intends to make the concept of mutual orientation relevant to the sociology of collaboration and interdisciplinarity (Jacobs and Frickel, 2009) and the social studies of algorithms.

Empirical context: the national strategy for AI development

In September 2020, the NSF announced the creation of six National Artificial Intelligence Research Institutes. Each Institute received $20 million in funding to be dispersed over the next five years. The Institutes were established through the efforts and support of federal agencies (National Science Foundation, U.S. Department of Agriculture, National Institute of Food and Agriculture, U.S. Department of Homeland Security, Science & Technology Directorate, the U.S. Department of Transportation, Federal Highway Administration) and industry partners (Amazon, Google, IBM, Intel, Nvidia, Accenture). Together, the Institutes will form the backbone of the national AI strategy.

Each of the Institutes has a designated theme. The rationale for an institute dedicated to trust has been justified differently by the AI experts and the NSF. While the NSF is invested in promoting a cross-agency framework for the ethical design of AI, the AI experts working with the weather forecasters see the Institute as an opportunity to resolve the jurisdictional struggle stemming in part from the black box problem of algorithmic predictions—predictions which often go against forecasters' intuitions. Dr. Amy McGovern, a computer scientist from the University of Oklahoma, directs the Institute. McGovern leads a team of experts from the fields of ML, atmospheric and ocean science, meteorology, and computer science. The Institute's secondary goal is to forge collaborations between academia, industry, and the private sector.[1]

First mutual orientation: Funding body and an alternative expertise network

The formation of the Institute

The Institute's origins can be traced to a pre-existing expertise network of computer scientists, weather and environmental scientists, and risk communication scholars. The key focus of this scholarly network was research on so-called "use-inspired" (or applied) ML. Institute director Amy McGovern explained that when the NSF released the call for proposals for National AI Institutes, she and her collaborators in meteorology had already begun to conceptualize a plan for a research institute focused on applying AI to atmospheric sciences. Some of the Institute's members have a substantial history of being funded by the National Oceanic and Atmospheric Agency's (NOAA) Joint Technology Transfer Initiative (JTTI). Thanks to this funding, even before the establishment of the Institute, its members dominated the research field in improving the automation of weather prediction.[2] In other words, there was already trust between the Institute's members and the weather prediction community.

Harry Collins et al. (2010) observe that initial trading zones, if successful, often culminate in a shared research proposal. This was the case with the alternative expertise network, arguably with Amy McGovern as one of its leaders. The centerpiece of the proposal was research on trust and AI. The following quote from McGovern illustrates that the AI and weather prediction experts understood the need to establish a common definition of trust:

> We need to work with our targeted set of end-users to learn how they're defining trustworthiness because it seems to be very different [from our definition].

With the release of the call for proposals, the team had to tailor the scope of their work and match definitions to the NSF's vision. While the Institute's focus on trust was prompted by the NSF's desire to establish a center for fundamental research questions on epistemological dimensions of trust in ML, the Institute's mission also became to alleviate the fear of meteorologists of being replaced by AI. The creation of the Institute was an effect of mutual orientation of a bottom-up vision generated by an alternative expertise network (Eyal, 2013) and the top-down framework for "Trustworthy AI" embraced by a federal funding agency, the NSF.

The crucial step in the mutual orientation between the NSF and the alternative expertise network warrants further explanation. While the NSF solicited proposals in the domain of trustworthy AI, the agency did not envision funding research in trustworthy AI, specifically in environmental sciences. The "Trustworthy AI Institute" could as likely focus on biomedicine or any other socially relevant domain. In other words, the NSF chose to orient its vision of future work on trustworthy AI towards a specific expertise network of ML experts already collaborating with weather forecasters and risk communication scholars.[3] As previous research shows, most often, "norms of AI are dynamic and are pieced together from various sources in traditional and transitional ways" (Gasser and Schmitt, 2020: 144). Likewise, the members of the Institute do not simply inherit the categories from the NSF Trustworthy AI framework; rather, they tailor the ethics at the Institute to suit their own experiences as well as the needs of the weather forecasting and the broader environmental science communities. As a result, the trustworthy AI framework became a boundary object, which enabled the initial expertise network to synthesize the definition of the funding agency, domain experts, and their views about what makes algorithmic models reliable.

I ground the discussion of trust in the definition adapted at the Institute from Meyer et al., (1995), which claims that trust is "the willingness to assume risk by relying on or believing in the actions of another party." I further discuss the definition of trustworthiness developed by the NSF.[4]

By making a distinction between a "relational" (involving relationships between actors) character of trust and "evaluative" (emphasizing the process of evaluation of claims, tools, or parties) character of trustworthiness, members of the Institute define "trustworthiness" as "a trustor's evaluation, or perception, of whether, when, why, or to what degree someone or something should or should not be trusted." These two definitions frame the internal work at the Institute.

Altogether, the Institute's work aims to bring together federal, industry, and professional standards of weather forecasting to engender a multidisciplinary workflow on developing trustworthy AI. The key component of overcoming both the fears of automation and mistrust in AI within a new expertise network are three related tasks: incorporating internal to the profession of weather forecasting standards of epistemic reliability, increasing model explainability, and aligning with social and environmental values of earth sciences.

Disrupting the man-machine mix in weather prediction

The institute's drive towards further automation in forecasting has the potential to impact the current dynamic in the long-standing history of the 'man-machine mix' (Henderson, 2017) mix in meteorology. In August 2021, McGovern became Editor-in-Chief of the most recent journal introduced by the American Meteorological Society called *Artificial Intelligence for the Earth Systems*. In a comment about the release of the journal, the president of the AMS, Michael Farrar explained:

> Artificial Intelligence and machine learning offer exciting opportunities to improve our understanding of weather, water, and climate. AMS is excited to host a new journal to improve the science of AI and its applications for AMS-related sciences.

The enthusiasm of the AMS about AI could be explained by the explicit work of the Institute towards establishing trust in the new technology and the legitimation of the novel epistemological model.

The introduction of a new technology into a professional domain engenders both fear and optimism. The enthusiasm of experts such as Michael Farrar about the inclusion of AI experts into their network of expertise—which mirrors the sentiment of many practitioners in the field—fits neatly within two key theoretical concepts originating in the sociological analysis of expertise introduced by Gil Eyal (2013), namely 'generosity' and 'co-production.' Drawing on the actor-network theory, Eyal describes generosity as being "the opposite of monopoly, namely, that a network of expertise, as distinct from the experts, becomes more powerful and influential by virtue of its capacity to craft and package its concepts, its discourse, its modes of seeing, doing, and judging, so they can be grafted onto what others are doing, thus linking them to the network and eliciting their cooperation" (Eyal, 2013: 875).[5] Eyal understands co-production as a process through which "a network of expertise becomes more powerful and influential by virtue of involving multiple parties—including clients and patients—in shaping the aims and development of expert knowledge" (Eyal, 2013: 876). The two concepts are meant to capture how "power consists not in restriction and exclusion, but an extension and linking" (Eyal, 2013: 876). In effect, the generosity and co-production help explain why the weather prediction community perceives AI as part of the strategy for establishing a more powerful expertise network and how AI experts seek to expand their methods into a new, socially relevant problem. In this context, the processes of mutual orientations could be seen as co-production and generosity at work.

"Trustworthy AI" framework and the National Science Foundation

The trustworthy AI framework, as defined by the NSF, bears the mark of its particular intellectual and organizational history. According to the NSF, a trustworthy AI should:

1. Be reliable;
2. Be explainable;
3. Adhere to privacy standards;
4. "Not exhibit biases that are socially harmful."

Using Mittelstadt et al.'s (2016) framework, we can distinguish that while the first two points could be categorized as epistemic concerns, the latter two points refer to normative concerns. This framework has its own history. Jeannette M. Wing[6] (2020) traces the history of conversations about trustworthy computing to the "Trust in Cyberspace" 1999 report by the National Research Council (1999). NSF joined this conversation in 2001 by initiating the Trusted Computing program in 2001 and later by expanding it through the Cyber Trust (2004), Trustworthy Computing (2007), and Secure and Trustworthy Systems (2011) programs (Wing, 2020). Wing observes that the industry soon followed the lead and began producing its own statements, beginning with Bill Gates' 2002 "Trustworthy Computing" memo (Gates, 2002). Some of the early reasons for articulating trust in digital technology had to do with the realization that cyberspace has become, towards the end of the 20th century, a critical national infrastructure prone to both attacks and disasters.[7] Defining what trust in digital infrastructures implies has been an area of discussion and ambiguity since that time. For example, the National Research Council report reads:

> The alert reader will have noted that the volume's title, Trust in Cyberspace, admits two interpretations. This ambiguity was intentional. Parse "trust" as a noun (as in "confidence" or "reliance"), and the title succinctly describes the contents of the volume—technologies that help make networked information systems more trustworthy. Parse "trust" as a verb (as in "to believe"), and the title is an invitation to contemplate a future where networked information systems have become a safe place for conducting parts of our daily lives. Whether "trust" is being parsed as a noun or a verb, more research is key for trust in cyberspace. (National Research Council, 1999: viii).

The subsequent iterations of the definition of trust attempted to ameliorate this ambiguity but also respond to technological developments. Therefore, it is reasonable to expect that the trustworthy AI framework has played a vital role in shaping the mission of one of the NSF's institutes since the Foundation has been deeply invested in defining and promoting the principles of trustworthy computing for over 20 years. Thus, we can see a refinement of a previous ethical statement in accordance with the existing "moral background" of AI development and an increase in the "complexity" of computational systems. The establishment of the Institute hence belongs to the long tradition of redefining trust in digital technology by nation-level actors.

Orienting ethics at the Institute towards NSF's trustworthy framework

The necessity for ethical standards in predictive analytics for environmental science is not a uniformly recognized need. For example, during one of the meetings, McGovern mentioned a pushback against implementing ethical training for environmental science from one of her colleagues from the National Academies of Arts and Sciences:

> I am now on the National Academies Board of atmospheric science and climate, and we're putting together a summer school on AI for Earth System prediction. We had a debate via email this week on whether or not ethics should be a part of that, and I held firm that yes, ethics needed to be part of that. One of the other people on the email chain was holding firm that there was no need for ethics in AI for Earth Science prediction because there was no reason that AI needed to be ethical because there was no bias that would show up. It wasn't that they were advocating that ethics was bad, just that they didn't think that there was any bias in anything that we were doing to predict in Earth Science.

While the ML experts do perceive a need to explore the epistemic grounds of ML predictions, they do not see the normative values as relevant to the application of ML in environmental sciences. Despite this ambivalence, the Institute members agree that the ethics of AI could and should be applied to the design of AI for earth sciences. There are four foundational domains and activities which facilitate a common ethical ground for the Institute. These are 1) reliance on the NSF's trustworthy AI framework, 2) establishment of an Institute-specific code of ethics, 3) formal educational activities—and specifically the core course called "AI, Ethics, and Geoethics" designed and taught by Amy McGovern, 4) discussions of ethical principles during regular, Institute-wide meetings. I will briefly describe each of these activities.

As mentioned above, the NSF Trustworthy AI framework is derived from the principles of trustworthy computing. Drawing on Mittelstadt and colleagues' typology, we see that the framework combines epistemic (reliability, explainability) and normative (privacy, social harm) concerns. However, this framework alone is not specific enough to serve the situated needs of AI in environmental sciences. Therefore, the Institute's code of ethics was derived from the codes of ethics of the American Meteorological Society, the American Geophysical Union, the American Association of Artificial Intelligence, and Google's AI Principles. The confluence of distinct disciplinary and organizational paradigms gave rise to a unique set of ethical considerations. While some of the standards in the Institute's code outline general principles of scholarly conduct, worth mentioning are points 3, 4,5, and 6 of the code (McGovern et al., 2020):

3. Stewardship of the Earth:

1. Members have an ethical obligation to weigh the societal benefits of their research against the costs and risks to human and animal welfare, heritage sites, or other potential impacts on the environment and society.
2. Members also have an ethical obligation to limit their contributions to climate change.

4. Public Communication:

1. Members have an ethical obligation to foster public awareness and understanding of AI, computing, related technologies, and their consequences.

5. When creating AI systems, members will:

1. Ensure that the public good is the central concern during all professional computing work.
2. Give comprehensive and thorough evaluations of AI2ES AI algorithms and their impacts, including analysis of possible risks.
3. Recognize and take special care of AI systems that become integrated into the infrastructure of society.

6. Members will create AI systems that will:

1. Avoid harm.
2. Protect the Earth and its environment including human and animal welfare.
3. Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing.
4. Be fair and take action not to discriminate.
5. Respect privacy.
6. Honor confidentiality.
7. Avoid creating or reinforcing bias.
8. Uphold high standards of scientific excellence.

The above principles have guided director McGovern during the design of her course on "Ethics of AI and Geoethics." The course serves both the student body at the University of Oklahoma, the Institute, and is publicly available on the Institute's website. The course reviews topics relevant to AI design, such as bias, transparency, liability, and security, and issues of social responsibility and interdisciplinary dynamics. The emphasis on interdisciplinary communication and collaboration draws on the work of and William Newell and Douglas Luckie (2019) on *Pedagogy for Interdisciplinary Habits of Mind* as well as other seminal works from the field of interdisciplinary studies and the research from the field of team science. McGovern's course poster for the Spring of 2021 prominently features the cover of Ruha Benjamin's book *Race After Technology* (2019), which, as she told me, significantly impacted her.

The language of trust becomes a pidgin (Galison, 1997) through which the Institute operates. Risk communication comes into the picture as the discipline most associated with regulating trust, and hence, they acquire a privileged position in deliberately setting out to comprehend the various definitions of trustworthy AI. Yet, as the forthcoming discussion will show, each group understands trust in a slightly different way. As Collins et al. (2010: 14) suggest, In some cases, interactional expertise trading zones rely on trade managed not by experts from each group who develop an interactional expertise, but rather by third parties who can talk to all groups involved. At the Institute, that risk communication scholars are the "third party" people managing a trade without the necessity for developing of an interactional expertise by other researchers at the Institute. This position is partly enabled by the risk communication scholars' expertise in qualitative methods: as social scientists, they are assumed to know how to translate across epistemic cultures. This translation process is tied to the perception that the language of qualitative social science offers a bird's eye view of the Institute.

One of the goals of the risk communication group (designated as Focus 3: Foundational research in AI risk communication for environmental science hazards) is to "Develop principled methods of using [the group's] knowledge and modeling to inform the development of trustworthy AI approaches and content, and the provision of AI-based information to user groups for improved environmental decision making." This goal is tied to achieving a certain level of pidgin-based communication between various research groups at the Institute. This is done through Institute-coordinated training and communications. Lead risk communication PI Ebert-Uphoff suggested that,

> One way to break down institution and discipline/topics barriers is to have a regular talk series. These talks need to be short and simple at the beginning, so the barrier is relatively low for Institute members to follow, regardless of their research background. Collaboration ideas and actions will most likely develop out of these "101" talks naturally.

What Ebert-Uphoff prescribes for the Institute aligns with Galison's (2010) observation that trade often relies on 'thin interpretation.' According to Peter Galison, "[t]rade focuses on coordinated, local actions, enabled by the thinness of interpretation rather than the thickness of consensus" (Galison, 2010: 36). The Institute-wide meetings often rely on such 'thinness.'

Lastly, the ethics of AI is one of the central topics for the bi-weekly Institute-wide meetings. Worth recounting here was a presentation given by Ebert-Uphoff titled "Responsible Use of AI—What role can [the Institute] play?" One of her slides states, "If the [Institute] does not address Responsible Use of AI for the weather/climate community, who will?" Ebert-Uphoff thus sees the Institute as a "role model" for other communities implementing AI in environmental sciences. The responsible use of AI, according to the author, should include two long-term goals: "Develop new techniques, customized for meteorology," and "Collect and translate existing solutions from [computer science] and other literature." During her presentation, Ebert-Uphoff drew attention to the concept of 'environmental injustice,' a process which she described this way:

> Due to limitations of sensors or other data sources, certain regions or certain meteorological conditions are under-represented in data. ML model learns from data; those scenarios are then under-represented in the ML model as well, which can quickly result in environmental injustice. … Air pollution and other sensors are more prevalent in affluent areas/countries. [and] Southern hemisphere often under-represented.

In response to this point, one of the attendees recounted an anecdote of someone who trained ML model to understand cyclones on data from the North without considering that on the Southern hemisphere, due to the Coriolis effect cyclones spin in the opposite direction. Furthermore, to show that "Using [neural networks] as a black box is not a good idea," in the same talk, Ebert-Uphoff used the story of Clever Hans, a horse who during the early 20th century was believed to have learned arithmetic. Clever Hans, as it turned out, was merely reading the cues of his trainer.[8]

The history of mistrust in automation in weather prediction

The emphasis on trust at the Institute intersects with long-standing tensions between computer modeling and the tacit expertise of weather forecasters—a tension between external and internal forces that came to define meteorology. As the history of weather prediction tells us, fears of automation are hardly new in this profession. This history also demonstrates that automation is not just an inevitable evolution but that it is led by experts from other domains—i.e., computer scientists, data scientists, AI experts.

A term that succinctly captures this professional tension is 'meteorological cancer'. Jennifer Henderson (2017) introduces this term in her ethnography of ethical dimensions of weather prediction. Henderson heard about this term from her interlocutors, who worried that younger forecasters, instead of "developing their own conceptual model" (Henderson, 2017: 1), use almost exclusively computer models to generate their forecasts. As meteorologists with whom Henderson (2017: 1) worked affirm, "[f]orecasters are substituting the computer model for their own knowledge." Henderson shows that the metaphor of 'meteorological cancer' captures the forecasters' realization that by downplaying the importance of their tacit expertise, they "are contributing to their own demise" (Henderson, 2017: 1). As with other professions, forecasters have for a long time been aware of their own, often elusive, position within the 'man-machine mix' (Henderson, 2017). Part of Henderson's (2017: 3) ethnographic goal was to understand the "competition of forecasters rivaling computer models for daily work even as the machines increasingly outperform them". This ethnographic account thus shows in detail how the fear of being automated out of a job manifests. Yet, the 'ontological fears' of weather forecasters, as Henderson calls them, are "not so much the loss of labor but the change in the image of themselves" (Henderson, 2017: 46).

The advent of modern weather forecasting is marked by the development of Numerical Weather Prediction (NWP) in the 1930s and 40s and the employment of computers to model atmospheric data. In Kristine Harper's words, the meteorologists sought to invest their energy and resources in developing NWP "to increase the fortunes of a research community that had long been on the margins of U.S. science" and, consequently, "to replace the art of forecasting with the science of meteorology" (Harper, 2012: 668). The meteorologists' goal, Harper (2012) argues, was to elevate meteorology to the status of a 'legitimate' and objective scientific discipline by increasing the quantitative element of the field.

Harper (2012) observes that there are two parallel views within the historiography of meteorology about who was more instrumental in shaping the field. One part of this literature emphasizes external actors, such as the polymath John von Neuman, who was deeply involved in designing the first computing system for weather forecasting. At the same time, other scholars attribute more substantial agentive capacity to meteorologists in defining their future. This argument aside, the point is that the birth of modern weather prediction is tied to the shift in the network of expertise in meteorology: NWP, in Harper's account, has been made possible by the "availability of a new and larger pool of scientifically educated and mathematically savvy meteorologists" (Harper, 2012: 670-71). Considering this history, I suggest that the mistrust of AI may result from not only a fear of 'technological unemployment' but also of destabilization of a professional identity. Furthermore, there is nothing new about the mistrust of automation, but the method of addressing it—by creating an intermediary professional organization—is a novel development.

As such, meteorology is one among a growing number of professions that face existential angst due to advancements in AI. Sociologist Phaedra Daipha (2015: 106) understands weather forecasting "as the art of collage" By this, Daipha means that weather forecasting is characterized by the "art of improvisation," or an ability to mobilize various streams of data and modeling and be competent in screenwork analysis, as well as actual observation of physical weather. Following Daipha and other ethnographers of weather professionals, I frame the introduction of AI into the field as part of the larger 'collage,' or "a heuristic that frames meteorological decision-making as a process of assembling, appropriating, superimposing, juxtaposing, and blurring of information" (Daipha, 2015: 21). Daipha further describes weather forecasting as 'art and science,' and foregrounds the blurring of the boundaries between human and the machine in the profession. In another register, what takes place at the Institute is a jurisdictional struggle between groups of experts who embrace 'mechanical objectivity' on the one hand and 'trained judgment' (Daston and Galison, 2010) on the other.

Second mutual orientation: Weather forecasters and machine learning experts

The introduction of AI in weather forecasting is a story of ethical and epistemological progression towards ever-increasing speed and accuracy of predictions. But what will make AI-based predictions more trustworthy? And "Who possesses the better understanding of the atmosphere: those who crunch the numbers, but never look outside, or those who are unimpressed by equations, but read the sky?" (Henderson, 2017: 689). Forecasters have asked this question for almost 70 years. In the story of Jay Forrester and the Whirlwind, Paul Edwards (1996) notes that it was Forrester who deliberately influenced high-ranking officials in the Office of the Naval Research, who initially were skeptical of the utility of the digital computer. Some generals were hostile to the idea that a machine could perform the tacit knowledge of strategizing. Analogously, the weather forecasting community has been characterized by friction between those who 'read the sky' and those who 'crunch the numbers,' to use Harper's (2012) words.

To gain the community's trust and secure a mutual orientation between ML and weather forecasting experts, the members of the Institute had to learn to do both. The ML experts at the Institute have understood the need to design ML that weather forecasters could trust. This form of ML is based on two central properties: explainability and adherence to the laws of physics. In sum, ML experts realized that to get the forecasters to trust their system, they needed AI to satisfy a number of criteria: 1) be explainable, 2) adhere to the laws of physics and look 'realistic,'[9] 4) adhere to the tacit norm of "erring on the side of caution." The following subsection examines these related criteria in AI design.

Trustworthy AI needs to be both explainable and realistic

McGovern has a long history of spanning the boundaries of computer science and weather forecasting. As a result, she has a unique vantage point to understand the role of explainability as a crucial factor in promoting ML for weather forecasting. McGovern's doctoral work was in computer science and on a type of AI called reinforcement learning (RL). At the University of Oklahoma, she holds a full professorship in both the School of Computer Science and the School of Meteorology. With expertise in ML and weather forecasting, she is a boundary-spanning figure (Aldrich and Herker, 1977; Ribes et al., 2019) who strives to present ML methods in ways the meteorology community can understand and trust. In a paper titled "Making the Black Box More Transparent: Understanding the Physical Implications of ML" (McGovern et al., 2019) published in *the Bulletin of the American Meteorological Society*, McGovern and colleagues argued:

> Despite its wide adoption in meteorology, ML is often criticized by forecasters and other end users as being a "black box" because of the perceived inability to understand how ML makes its predictions. This phenomenon is not exclusive to meteorology, and many ML practitioners and users have recently begun to focus on this interpretability problem (McGovern et al., 2019: 2176).

This problem statement made by a computer scientist in a prime venue for meteorological research attests to the gravity of the problem of trust in automation in weather forecasting.

But professional meteorologists are not the only group the Institute's members seek trust from. In one of the talks about the Institute presented at the "2nd Workshop on Leveraging AI in Environmental Sciences" organized by NOAA, McGovern said: "You can't develop one particular AI technique that's going to meet all of these needs. What you need to do is to take into account the end user's needs and to make it trustworthy for that end user—you need to care about the end-user that you're looking at." This approach to trustworthiness requires deliberate tailoring of both algorithmic tools and discourses that explain these tools to different publics and actors.

According to the Institute members, the key to solving the problem of trust in AI among both the forecasters and the public is explainability. In one of the lectures she delivered about the Institute, McGovern described the mission of the Institute in the following manner:

> we're working on developing explainable AI methods that are aligned with environmental science, domain perspectives and priorities. This means that we care about what environmental scientists care about. We care about the spatial and temporal nature of the data. We care about the physics-based nature of the data, etcetera. So, it isn't just an explainable AI method that's developed theoretically—we're testing it on all the environmental science domains.

During one of the Institute meetings, McGovern stated succinctly:

> what is the future of everything we are trying to do? I think we need to integrate the AI and the physics and the robust approaches that we've started with explainable AI.

McGovern and her team are devoted to a solution-oriented approach grounded in problems emerging from environmental sciences. For the Institute members, specificity matters—the spatial, temporal, and physical aspects of environmental data, as well as the needs of potential users, need to be considered. As one researcher at the Institute said, "If we're going to be showing the results of these [predictive models] to [forecasters] to say: 'this is trustworthy,' you can't give them stuff that doesn't look realistic at all."

Not all AI models have laws of physics built into them. And for weather forecasters, physics-based AI is one of the conditions for the technology's reliability and realism. "Physics-based AI" is a family of AI models that respects actual physical processes, such as the dynamics of storm formation. Making ML models physics-based is part of the process of establishing trust in the model, specifically in the domain of weather forecasting. In effect, the ML experts themselves have to develop an understanding of the physics of weather. This is one of the most challenging things to train as an ML expert. Forecasters need translators who can explain how algorithms arrive at their results. While prior to the establishment of the Institute, this translation had to be managed solely between AI and weather experts, the Institute adds an extra layer of risk communication scholars who, through their social science sensitivity, can help mediate across epistemic cultures (Knorr-Cetina, 1999).

In this context of interdisciplinary translation work, I want to draw attention to the complexity of 'opening the black box' of algorithms as a solution to the problem of trust. As Anthony Giddens argued,

> There would be no need to trust anyone whose activities were continually visible and whose thought processes were transparent or to trust any system whose workings were wholly known and understood. (Giddens, 1990: 33)

Following Giddens, we can conclude that full explainability would ostensibly make the articulation of standards of trust obsolete. But full explainability is rarely attainable. Trust requires more than just explainability.

Analyses of explainability and transparency have been a key trope among critical algorithmic and data studies scholars, some of whom have observed limitations of the notion of transparency. For example, Mike Ananny and Kate Crawford (2018: 5) observe that transparency "assumes that knowing is possible by seeing, and that seemingly objective computational technologies like algorithms enact and can be held accountable to a correspondence theory of truth". But as Ananny and Crawford make apparent, transparency does not necessarily build trust. On the other hand, Cynthia Rudin and Joanna Radin (2019) questioned whether we need to make 'black-boxed' AI in the first place. They argued that "an accurate machine or an understandable human" (Rudin and Radin, 2019: 4) is a false dichotomy. The explainability of AI systems is undeniably a virtue that researchers strive for, but as Rudin and Radin point out, the dichotomy between "Being asked to choose an accurate machine or an understandable human is a false dichotomy" (Rudin and Radin, 2019: 4). While arguably the philosophy of technology at the Institute reproduces this dichotomy, the question remains how the Institute will help to resolve this tension, and how explainability will influence the uptake of AI in weather prediction.

When machine learning goes out to lunch and predicts the end of the world: Calibration and mutual orientations at the NOAA's Storm Prediction Center

The fundamental conundrum in the forecaster's work is to determine whether to trust automated prediction generated by a computer or their own intuitions. The mode of prediction in ML clashes with institutional norms and forecasters' intuitions. As Henderson (2017) notes, a part of forecasters' trust is based on reducing their exposure to criticism by 'under forecasting'—meaning here simply to communicate to the public lower probabilities than those generated by their mental and digital models. Yet, ML models do not hold to this facet of an 'ethic of accuracy' of weather prediction (Henderson, 2017; see also MacKenzie, 1987). Paramount in this context is the importance of "calibration" between forecasters' predictions and the predictions of ML models, which, while not as accurate in the eyes of ML experts as it could be, respects the implicit norms of weather forecasters. By calibration, here I mean a process of translating ML models into more realistic forms of prediction. The next few paragraphs offer an example of a tension between ML researchers who try to be as accurate as possible and forecasters who lean towards performing cautious predictions.

During one of the Institute meetings, director McGovern recollected an event during which a model deemed efficient by the ML experts was considered to be untrustworthy by the NOAA's Storm Prediction Center (SPC) forecasters. The point of contention was a divergence between the ML models' and the forecasters' predictions. One Institute member explained: "on this particular day, the ML model gave 80% [probability] of a certain temperature, while the SPC issued the probability of 50%." Critically, due to the tacitly accepted principle of "erring on the side of caution" (Henderson, 2017: xxxvi), the highest probability forecasters wanted to issue was 60%.[10] McGovern elaborated: "From our perspective, a 100% probability wasn't a problem (…) if the model says a 100, why shouldn't we say a 100? But the SPC said 'hell no.'" Another ML expert remarked: "You can see that by design, SPC wants to under-forecast." The same expert put this divergence in the context of their long-term work with the SPC:

> You're trying to defend the model the first year. And [the SPC people] would just flip past because it's like: 'Oh yeah, the ML is out to lunch again, it's putting 80%.' (…) So, when they're looking at 30% as a high-end event, and a model is putting out 85%, they're looking at it and saying, 'Oh, this model is basically predicting the apocalypse for every day, and we can't trust it if it's always predicting the end of the world.'

The forecasters could not trust the ML models because they suggested probabilities much higher than they were used to. But, as this particular ML researcher explained, "By default, the model doesn't have any sort of idea of what SPC predicts. It just gives you a raw weighting based on (…) the data." To remedy this divergence, designing physics-based AI and calibration of models has emerged as a critical issue. The concept of calibration in this scenario becomes one of the modes of a mutual orientation between two groups of experts. To make AI probabilities align more with forecasters' norms, AI researchers calibrated the model which previously "has gone to lunch" with multiple real-life datasets. Only after the appropriate calibration took place were the forecasters' and models' predictions aligned.

This scenario exemplifies a classical problem of expert system approach to AI and the longstanding tension between predictive experts and computers. One of the solutions to the calibration problem in the eyes of the Institute members is to extract 'mental models' of forecasting and input them into ML predictions. Imme Ebert-Uphoff mentioned that it would be beneficial to use social science methods to understand how forecasters read the data and build AI based on those 'mental models.' She emphasized "getting feedback from social scientists about how we should develop ML methods" and explained that she does "a little bit of interviewing" when she sits down with an end-user and asks: "how do you do [predictive work] right now?" In her experience, forecasters often clash with the computer science people who say, "let the computer do it all." In one of the talks, she concludes:

> We could do the whole community a big favor by revisiting the entire topic—not just what explanations should be like, but can we make

> ML a little bit more like what people do manually right now? (…) Can we build a mechanism and vocabulary where we can actually talk about it?

A fascinating aspect of this situated process of calibration is that despite a historical decline in the prominence of the expert system approach to AI development, the ethics of explainability pushes some AI developers to once again revisit this less prominent form of AI[11] In recent years, it was the data-driven system that won the battle, but the Institute is evincing the prowess of the expert systems approach.

Discussion

This case study testifies to the necessity of supplementing contemporary critique of AI with historical analysis. Initiatives like the Cambridge University seminar series on "Histories of Artificial Intelligence: A Genealogy of Power" are among many scholarly developments promoting an integrative, historical, and sociological examination of AI. From such a vantage point, the introduction of AI into weather forecasting can be better understood within the context of a *longue durée* of the interplay between trained judgment and mechanical objectivity in weather prediction (Daston and Galison, 2010). For example, Henderson argues that the crux of the matter is that:

> Amid the talk of competition between humans and their technologies, then emerges a tension between the success of their work as predictive experts, which computer models help facilitate, and the value of their own expert skill in the process. At stake are the identities of forecasters as scientists and the survival of their profession in ways they envision it ought to exist (Henderson, 2017: 10)

She adds: "In a forecasting office, boundaries between human and computer are fluid, blurred, and multiple. There is no single human nor a solitary machine but a plurality of both" (Henderson, 2017: 10). As with other professions, AI might merely reposition the boundaries between the human and the computer. Nonetheless, such repositioning can turn into a professional 'identity crisis' (Henderson, 2017: 11). The possible disruption in the professional identity of forecasters caused by AI lies at the core of the jurisdictional struggle explored in this paper. Interpreting such emerging jurisdictional struggles with attention to the history of automation might prove to be analytically advantageous for STS scholars, as well as for technology designers and policymakers.

The establishment of the Institute suggests that the introduction of AI into domains such as weather forecasting and environmental sciences at large necessitates deliberate training in novel, hybrid forms of expertise. The need for calibration between the professional norm of "under forecasting" and AI's predictions also illustrates the tacit dimensions of professional expertise. This state of affairs is present in other professional contexts as well. For example, an ethnographic study of predictive policing revealed the importance of forming new intermediary occupational roles as a key to securing trust in AI (Waardenburg et al., 2018). Such professional intermediaries helped to establish "the superiority of algorithmic decisions over human expertise," but their presence also "further black-box[ed] the inherent inclusion of human expertise" in making decisions based on AI reccomendations (Waardenburg et al., 2018: 14). The capacity to interpret 'black-boxes' comes with a specific form of intellectual capital. Those possessing such expertise might likely succeed in gaining prominence within the larger domain. Therefore, the introduction of professional intermediaries might have profound effects on the future of environmental prediction.

Relatedly, as a growing literature focused on environmental data practices alerts us to the unique characteristics of environmental data (e.g., Fortun et al., 2016; Gabrys, 2016, 2020; Lippert, 2015), critical algorithm and data scholars will need to pay more attention to the formulation of AI and data ethics in environmental sciences. Recounted above debate among the Institute members about whether AI in Earth Sciences might exhibit biases is a case in point. To repeat, as McGovern put it, one of her collegues argued that "there was no need for ethics in AI for Earth Science prediction because there was no reason that AI needed to be ethical because there was no bias that would show up." Arguably, environmental STS analysis can offer a more nuanced and situated view of algorithmic bias and ethics at large. Interdisciplinary communication across social, environmental, and computer sciences is becoming more ubiquitous in AI design, and the Institute might offer many best practices for such collaborations. On the STS side, many scholars have adopted an openly collaborative ethos, as for example in Gina Neff and collegues' "practice-based framework for imporving critical data studies and data science" (Neff et al., 2017: 85), and such frameworks might also prove generative for studies of both data and algorithms and for fostering interdisciplinary dialogue.

The study of the orientations of scientific research toward socially relevant problems has produced many insights into the formation of new scientific movements and disciplines (Frickel, 2004; Jacobs, 2014; Hess et al., 2008). Nevertheless, there is a pressing need for further research about the role the ethics of technology design plays in the formation of networks of expertise consisting of private, public, and academic actors. Ethics statements common to industry often have objectives distinct from those embraced by public agencies or universities, and how the many genres of ethical frameworks are consolidated will require further study. Multi-sector collaborations engender the composition of unique and discipline-tailored ethical standards, thus putting into question the utility of "one size fits all" design standards.

Multi-sector organizations might prove to be very effective spaces for translating hig level policy discourses and moral backgrounds of technology development for the purpose of discipline-specific use of AI. Paul Edwards (1996) argued that the funding of Forrester's Whirlwind project—a project that paved the way for semi-automatic command and control systems in the army to the dismay of many high-ranking officials—could not be possible outside of the political milieu of the late 1940s and early 1950s. Analogously, the contemporary cultural conversation and policy discourses about the ethics of AI were a causal factor in instigating a mutual orientation between the NSF, the Institute, and weather forecasters. But the middle ground between policy discourses and technology development is often occopuied by boundary organizations (Guston, 1999; Vaughan, 1999). More research is necessary to capture how situated and idiosyncratic standards of AI design become stabilized and embraced by muti-sector projects and organizations, especially as more private-public partnerships (such as US NSF AI Institutes) are created under the often-seemingly over-arching umbrella of national AI strategies.

Conclusions

The evincing of ethical and socially desirable image plays a significant role during the emergence and institutionalization of alternative expertise networks. This is often done, as in the case of the Institute, through an alignment with pre-existing ethical standards or moral backgrounds (Abend, 2014) of technology design, but it also involves extensive, domain-specific adjustment of standards. This study intervenes in the literature on the ethics of AI by showing that the prevailing moral background and national ethical standards of technology development, while crucial, are by themselves insufficient in providing tailored solutions to domain-specific issues associated with trust in new technologies. The use of the concept of 'mutual orientations' and the reading of the literature on the sociology of expertise and SCOT approach offers an analytical purchase on the question of how alignments of design standards shape emerging expertise networks and the introduction of AI into an existing predictive science. Moreover, the concept of 'mutual orientations' highlights the dynamic nature of the institutionalization of hybrid expertise networks. Indeed, introducing new technology into a professional field is often led by a desire to form a more robust network through generosity and co-production (Eyal, 2013). The above analysis highlights the importance of not only the role of transparency in forging trust between experts but also of a design process sensitive to the norms and standards of an expert community. For the experts at the Institute, this meant creating algorithms that adhered to the laws of physics and intuitions and norms of weather forecasters.

References

Abbott A (2014) *The System of Professions: An Essay on the Division of Expert Labor*. Chicago: University of Chicago press.

Abend G (2014) *The Moral Background*. New Jersey: Princeton University Press.

Aldrich H and Herker D (1977) Boundary Spanning Roles and Organization Structure. *Academy of Management Review* 2(2): 217–230.

Ananny M and Crawford K (2018) Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability. *New Media & Society* 20(3): 973–989.

Benanav A (2020) *Automation and the Future of Work*. London: Verso.

Benjamin R (2019) *Race After Technology: Abolitionist Tools for the New Jim Code*. Hoboken: John Wiley & Sons.

Besteman C and Gusterson H (2019) *Life by Algorithms: How Roboprocesses Are Remaking Our World*. Chicago: University of Chicago Press.

Bijker W E (1997) *Of Bicycles, Bakelites, and Bulbs: Toward a Theory of Sociotechnical Change*. Cambrdige: MIT Press.

Brady J (2018) Toward a Critical, Feminist Sociology of Expertise. *Journal of Professions and Organization* 5(2): 123–38.

Brysse K, Oreskes N, O'reilly J and Oppenheimer M (2013) Climate change prediction: Erring on the side of least drama? *Global Environmental Change* 23(1): 327-337.

Callon M (1984) Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay. *The Sociological Review* 32(1_suppl): 196–233.

Christin A (2020) The Ethnographer and the Algorithm: Beyond the Black Box. *Theory and Society* 49(5): 897–918.

Collins H, Evans R and Gorman M (2007) Trading Zones and Interactional Expertise. *Studies in History and Philosophy of Science Part A* 38(4): 657–666.

Collins H, Evans R and Gorman M (2010) Trading Zones and Interactional Expertise. In: Gorman M (ed) *Trading Zones and Interactional Expertise: Creating New Kinds of Collaboration*. Cambridge: MIT Press, pp.7-24.

Crawford K (2021) *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.

Daipha P (2015) *Masters of Uncertainty*. Chicago: University of Chicago Press.

Daston L and Galison P (2010) *Objectivity*. New Jersey: Princeton University Press.

d'Aquin M, Troullinou P, O'Connor N E, Cullen A, Faller G and Holden L (2018) Towards an ethics by design methodology for AI research projects. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 54-59.

Dick S (2019) Artificial Intelligence. *Harvard Data Science Review* 1(1).

Dubber M D, Pasquale F and Das S (2020) *The Oxford Handbook of Ethics of AI*. Oxford: Oxford Handbooks.

Edwards P N (1996) *The Closed World: Computers and the Politics of Discourse in Cold War America*. Cambridge: MIT Press.

Epstein S (1995) The Construction of Lay Expertise: AIDS Activism and the Forging of Credibility in the Reform of Clinical Trials. *Science, Technology, & Human Values* 20(4): 408–37.

Epstein S (2008) *Inclusion: The Politics of Difference in Medical Research*. Chicago: University of Chicago Press.

Eubanks V (2018) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Eyal G (2013) For a Sociology of Expertise: The Social Origins of the Autism Epidemic. *American Journal of Sociology* 118(4): 863–907.

Fleagle RG (1986) NOAA's Role and the National Interest. *Science, Technology, & Human Values* 11(2): 51–62.

Floridi L (2014) Technological Unemployment, Leisure Occupation, and the Human Project. *Philosophy & Technology* 27(2): 143–50.

Ford M (2015) The Rise of the Robots: Technology and the Threat of Mass Unemployment. *International Journal of HRD Practice Policy and Research* 111.

Fortun K, Poirier L, Morgan A, Costelloe-Kuehn B and Fortun M (2016) Pushback: Critical Data Designers and Pollution Politics. *Big Data & Society* 3(2): 2053951716668903.

Frickel S (2004) *Chemical Consequences: Environmental Mutagens, Scientist Activism, and the Rise of Genetic Toxicology*. New Brunswick: Rutgers University Press.

Frickel S and Gross N (2005) A General Theory of Scientific/Intellectual Movements. *American Sociological Review* 70(2): 204–232.

Gabrys J (2016) Practicing, Materialising and Contesting Environmental Data. *Big Data & Society* 3(2): 2053951716673391.

Gabrys J (2020) Smart Forests and Data Practices: From the Internet of Trees to Planetary Governance. *Big Data & Society* 7(1): 2053951720904871.

Galison P (1997) *Image and Logic: A Material Culture of Microphysics*. Chicago: University of Chicago Press.

Galison P (2010) Trading with the enemy. *Trading zones and interactional expertise: Creating new kinds of collaboration* 21(1): 147-175.

Gasser U and Schmitt C (2020) The Role of Professional Norms in the Governance of Artificial Intelligence. In: Dubber M D, Pasquale, F and Das S (eds) *The Oxford Handbook of Ethics of AI*. Oxford: Oxford University Press, pp. 141-162.

Gates B (2002) Trustworthy Computing Memo. Available at: https://news.hitb.org/content/complete-text-bill-gates-trustworthy-computing-memo (accessed 8.1.2023).

Giddens A (1990) *The Consequences of Modernity*. Redwood City: Stanford University Press.

Gieryn T F (1983) Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists. *American Sociological Review* 48(6): 781–795.

Greene D, Hoffmann A L and Stark L (2019) Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. In: *Proceedings of the 52nd Hawaii international conference on system sciences*.

Guston D H (1999) Stabilizing the Boundary between US Politics and Science: The Role of the Office of Technology Transfer as a Boundary Organization. *Social Studies of Science* 29(1): 87–111.

Hagendorff T (2020) The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines* 30(1): 99–120.

Harper K C (2012) *Weather by the Numbers: The Genesis of Modern Meteorology*. Cambrdige: MIT Press.

Henderson J J (2017) To Err on the Side of Caution: Ethical Dimensions of the National Weather Service Warning Process. Dissertation. Virginia Tech.

Hess D, Breyman S, Campbell N and Martin B (2008) Science, Technology, and Social Movements. In: Hackett E J, Amsterdamska O, Lynch ME and Wajcman J (eds) *The Handbook of Science and Technology Studies.* Cambridge: MIT Press, pp. 473-498.

Hoffman R R, Mueller S T, Klein G and Litman J (2018) Metrics for Explainable AI: Challenges and Prospects. *ArXiv Preprint ArXiv:1812.04608*.

Iliadis A and Russo F (2016) Critical Data Studies: An Introduction. *Big Data & Society* 3(2): 2053951716674238.

Jacobs J A and Frickel S (2009) Interdisciplinarity: A critical assessment. *Annual review of Sociology* 35: 43–65.

Jacobs J A (2014) *In Defense of Disciplines*. Chicago:University of Chicago Press.

Keynes J M (1930) Economic Possibilities for Our Grandchildren; Scanned by Yale University Economics Department from John Maynard Keynes, Essays in Persuasion. New York: Norton.

Kitchin R (2014) Big Data, New Epistemologies and Paradigm Shifts. *Big Data & Society* 1(1): 2053951714528481.

Kline R R (2015) *The Cybernetics Moment: Or Why We Call Our Age the Information Age*. Baltimore: Johns Hopkins University Press.

Knorr-Cetina K (1999) *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge: Harvard University Press.

Lippert I (2015) Environment as Datascape: Enacting Emission Realities in Corporate Carbon Accounting. *Geoforum* 66: 126–35.

Lipton Z C and Steinhardt J (2018) Troubling Trends in Machine Learning Scholarship. *ArXiv Preprint ArXiv:1807.03341*.

MacKenzie D (1987) Missile Accuracy: A Case Study in the Social Processes of Technological Change. In: Bijker W E, Hughes T P and Pinch T (eds) *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. Cambridge: MIT Press, pp. 195–222.

Mannheim K (2013) *Ideology and Utopia*. London: Routledge.

McClure P K (2018) "You're Fired,' Says the Robot: The Rise of Automation in the Workplace, Technophobes, and Fears of Unemployment.' *Social Science Computer Review* 36(2): 139–56.

McGovern A, Lagerquist A, Gagne R et al. (2019) Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning. *Bulletin of the American Meteorological Society* 100(11): 2175–99.

McGovern A, Neeman H, Hickey J and Gagne D J (2020) AI2ES Strategic and Implementation Plan. Available at: https://docs.google.com/presentation/d/15Xu2Ze1EBA0T3KESOEQeXL4ZRX66Y7xMocpJs2XHcSw/edit#slide=id.p (accessed: 01.07.2024).

Mendon-Plasek A (2021) Mechanized Significance and Machine Learning: Why It Became Thinkable and Preferable to Teach Machines to Judge the World. In: Roberge J and Castelle M (eds) *The Cultural Life of Machine Learning*. Heidelberg: Springer, pp. 31-78.

Metcalf J and Moss E (2019) Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics. *Social Research: An International Quarterly* 86(2): 449–76.

Mittelstadt B D, Allo P, Taddeo M, Wachter S and Floridi L (2016) The Ethics of Algorithms: Mapping the Debate. *Big Data & Society* 3(2): 2053951716679679

Moats D and Seaver N (2019) 'You Social Scientists Love Mind Games': Experimenting in the 'Divide' between Data Science and Critical Algorithm Studies. *Big Data & Society* 6(1): 2053951719833404.

National Academy of Engineering, U. S. (2004) *The Engineer of 2020: Visions of Engineering in the New Century*. Washington: National Academies Press.

National Research Council (1999) *Trust in Cyberspace*. Edited by Schneider FB, Washington: The National Academies Press.

Neff G, Tanweer A, Fiore-Gartland B and Osburn L (2017) Critique and contribute: A practice-based framework for improving critical data studies and data science. *Big data* 5(2): 85-97.

Newell W H and Luckie D B (2019) Pedagogy for Interdisciplinary Habits of Mind. *Journal of Interdisciplinary Studies in Education* 8(1): 6–20.

Nichols T (2017) *The Death of Expertise: The Campaign against Established Knowledge and Why It Matters*. Oxford: Oxford University Press.

Olhede S C and Wolfe P J (2018) The Growing Ubiquity of Algorithms in Society: Implications, Impacts and Innovations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376(2128).

O'Mahony S and Bechky B A (2008) Boundary Organizations: Enabling Collaboration among Unexpected Allies. *Administrative Science Quarterly* 53(3): 422–459.

Oreskes N (2019) *Why Trust Science?* New Jersey: Princeton University Press.

Oui J (2022) Commodifying a "good" weather data: commercial meteorology, low-cost stations, and the global scientific infrastructure. *Science, Technology, & Human Values* 47(1): 29-52.

Pasquale F (2015) *The Black Box Society*. Cambridge: Harvard University Press.

Pasquale F (2019) A Rule of Persons, Not Machines: The Limits of Legal Automation. *George Washington Law Review* 87(1).

Pinch T J and Bijker W E (1984) The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social studies of science* 14(3): 399–441.

Polanyi M (2009) *The Tacit Dimension*. Chicago: The University of Chicago Press.

Porter T M (2020) *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. New Jersey: Princeton University Press.

Powers T and Ganascia J-G (2020) *The Ethics of the Ethics of AI*. Oxford: Oxford University Press.

Randalls S (2010) Weather Profits: Weather Derivatives and the Commercialization of Meteorology. *Social Studies of Science* 40(5): 705–730.

Rabeharisoa V and Callon M (2002) The involvement of patients' associations in research. *International Social Science Journal* 54(171): 57–63.

Randalls S (2017) *Commercializing Environmental Data: Seeing like a Market.* London: Routledge.

Ryan M (2020) In AI we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics* 26(5): 2749–2767.

Ribes D, Hoffman A S, Slota S C and Bowker G C (2019) The Logic of Domains. *Social Studies of Science* 49(3): 281–309.

Richterich A (2018) *The Big Data Agenda: Data Ethics and Critical Data Studies*. London: University of Westminster Press.

Rottner R (2019) Working at the Boundary: Making Space for Innovation in a NASA Megaproject. *Social Studies of Science* 49(3): 403–431.

Rudin C and Radin J (2019) Why Are We Using Black Box Models in AI When We Don't Need to? A Lesson from an Explainable AI Competition. *Harvard Data Science Review* 1(2).

Shapin S (1994) *A Social History of Truth: Civility and science in seventeenth-century England*. Chicago: University of Chicago press.

Shapin S (2008) *The Scientific Life: A Moral History of a Late Modern Vocation*. Chicago: University of Chicago Press.

Slayton R and Clarke B (2020) Trusting Infrastructure: The Emergence of Computer Security Incident Response, 1989–2005. *Technology and Culture* 61(1): 173–206.

Star S L and Griesemer J R (1989) Institutional Ecology, Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science* 19(3): 387–420.

Vaughan D (1999) The Role of the Organization in the Production of Techno-Scientific Knowledge. *Social Studies of Science* 29(6): 913–943.

Waardenburg L, Sergeeva A and Huysman M (2018) Hotspots and Blind Spots. In: *Working Conference on Information Systems and Organizations*. Heidelberg: Springer, pp. 96-109.

Waidzunas T (2013) Intellectual Opportunity Structures and Science-Targeted Activism: Influence of the Ex-Gay Movement on the Science of Sexual Orientation. *Mobilization: An International Quarterly* 18(1): 1–18.

Wing J M (2020) Trustworthy Ai. *ArXiv Preprint ArXiv:2002.06276*.

Notes

1 On the commercialization of meteorology and weather data, see Randalls, 2010, 2017; and Oui, 2022; and on the effects of the business-oriented model on the history of NOAA, see Fleagle, 1986.

2 For an analogous analysis of a relationship between politics and science, see Guston's (1999) work on the Office of Technology Transfer as a 'boundary organization.'

3 Boundary organizations, like the Institute, often bring unexpected collaborators together (O'Mahony and Bechky, 2008).

4 Trust has been defined differently in other institutional (i.e., Mozilla Foundation) and political (European Union) contexts (see Greene et al., 2019).

5 Eyal borrows the concept of co-production not directly from the seminal work of Sheila Jasanoff but from Vololona Rabeharisoa and Michel Callon's (2002) reading of Jasanoff.

6 Wing, who is now the director of the Data Science Institute at Columbia University, has been deeply involved in shaping the NSF's perspective on trust as an Assistant Director of the Computer and Information Science and Engineering Directorate between 2007 and 2010.

7 For the historical context about trust and computing infrastructure, see Slayton and Clarke, 2020.

8 See Kate Crawford (2021: 4) on the myth of Clever Hans.

9 'Realistic,' a term forecasters use, means in this context that ML models need to respect the laws of physics. Making ML models physics-based is, hence, part of the process of establishing trust in the model.

10 Brysse et al. (2013) observed a similar bias towards 'erring on the side of least drama' among climate scientists who, contrary to some accusations of alarmism, often underpredict future climate change.

11 For the history of the relationship between expert systems and data-driven approaches in ML, see Dick, 2019 and Mendon-Plasek, 2021.