DOI: 10.1002/pra2.275

Good systems, bad data?: Interpretations of AI hype and failures

Stephen C. Slota[1] | Kenneth R. Fleischmann[1] | Sherri Greenberg[2] | Nitin Verma[1] | Brenna Cummings[2] | Lan Li[1] | Chris Shenefiel[3]

[1]School of Information, The University of Texas at Austin, Austin, Texas

[2]LBJ School of Public Affairs, The University of Texas at Austin, Austin, Texas

[3]Cisco Systems, San Jose, California

Correspondence
Stephen C. Slota, School of Information, University of Texas at Austin, 1616 Guadalupe St, Suite #5.202, Mail Code: D8600, Austin, Texas 78701-1213
Email: steveslota@gmail.com

Abstract

Artificial intelligence (AI), including machine learning (ML), is widely viewed as having substantial transformative potential across society, and novel implementations of these technologies promise new modes of living, working, and community engagement. Data and the algorithms that operate upon it thus operate under an expansive ethical valence, bearing consequence to both the development of these potentially transformative technologies and our understanding of how best to manage and support its impact. This paper reports upon an interview-driven study of stakeholders engaged with technology development, policy, and law relating to AI. Among our participating stakeholders, unexpected outcomes and flawed implementations of AI, especially those leading to negative social consequences, are often attributed to ill-structured, incomplete, or biased data, and the algorithms and interpretations that might produce negative social consequence are seen as neutrally representing the data, or otherwise blameless in that consequence. We propose a more complex infrastructural view of the tools, data, and operation of AI systems as necessary to the production of social good, and explore how representations of the successes and failures of these systems, even among experts, tend to valorize algorithmic analysis and locate fault at the quality of the data rather than the implementation of systems.

KEYWORDS

critical infrastructure studies, ethics of artificial intelligence, media hype, value-sensitive design

1 | INTRODUCTION

How does a good system go bad? Popular media contain a variety of accounts of the successes and failures of Artificial Intelligence (AI)-based systems, as well as a set of narratives as to how these systems produce either successes or failures. Image recognition systems are lauded for their ability to parse x-rays more effectively than trained doctors, and question-answering systems compete on Jeopardy. Alongside these successes are disturbing evidence of racist prison sentencing and medical prioritizing algorithms, and discriminatory hiring algorithms. In many of these cases, the hype of the system is placed at the algorithm, the ever black-boxed, unknowable learning system, where the failures are placed at the data as incomplete, ill-structured, or biased in its representation.

83rd Annual Meeting of the Association for Information Science & Technology October 25-29, 2020.

AI has developed over several decades and is now a prominent and mature field of scholarship (Stone et al., 2016). Symbolic AI focusing on the representation and use of domain knowledge was an early area of AI research (Ribes et al., 2019). Much recent innovation has been in the domain of statistical machine learning, including deep learning approaches using artificial neural networks, including applications in domains such as natural language processing, computer vision, and robotics (Stone et al., 2016). Modern AI is particularly reliant on engagement with large data sets, processed and weighted with some level of autonomy, and delivering probabilistic, rather than deterministic, results.

The ethical, policy, and legal concerns of AI are not yet well-defined. The variety and quantity of data as well as the often-counterintuitive outputs of algorithmically driven analysis make it significantly more difficult to predict harms. The data used to support AI is drawn from a large number of sources, including from people who may not even be aware that the data is being collected for that purpose. However, the conclusions derived from the application of AI to this heterogenous data often have the weight of knowledge without significant accounts of their uncertainty. One key challenge in this domain are issues arising from attempts to understand and assert the accountability of “black-boxed” analytic techniques (Fleischmann & Wallace, 2005, 2009)—particularly when the outcomes of research conducted using those techniques are used to inform policy, direct resources, and respond to emergency situations (Lehr & Ohm, 2017).

Popular accounts of AI successes and failures are not without discursive power. It is often difficult to discuss ‘bad’ AI without thinking of HAL, Skynet, or other similar media representations of AI gone wrong (in fact, these popular portrayals were often referenced by our interview subjects when they were asked about potential negative consequences of AI). Similarly, positive popular accounts of AI often fail to account for the flaws and limitations of these systems or otherwise do not transparently represent their operation or scope. Kranzberg's (1986) First Law of Technology holds, “technology is neither good nor bad, nor is it neutral” (p. 547). The extreme examples of overhyping the positive or negative implications of AI fall into the camps of viewing AI as purely good or bad. However, it is also important to note that AI is not neutral, some AI systems have some good implications and some bad implications for different members of a given society or across societies. Thus, the challenge comes in determining which factors influence the “goodness” or “badness” of AI.

One common scapegoat for bad AI is bad data. Modern AI is characterized by its relationship to, and reliance upon, broad, heterogeneous regimes of data collection and analysis. Algorithmic analysis provides tools to deal with and analyze increasingly broad and heterogeneous data sources towards a variety of potential societal outputs. The role of government, for example, in managing the deviations in stock prices emerging from the practice of algorithmic trading, particularly when that trading is international in scope (Brogaard, Hendershott & Riordan, 2017; Cartea, Donnelly, & Jaimungal, 2017; Chaboud & Chiquoine, 2014). Under the reality of imminent change—driven by technology but bearing substantive social effects and reflecting a changing conception of the possible—issues of governance and ethics of AI are becoming substantially more important, but not necessarily better articulated or understood (Barocas & boyd, 2017; Kitchin, 2014).

This paper reports findings from an interview study of stakeholders in the field of AI, including lawyers, legal scholars, policymakers, and government workers, as well as the designers and researchers directly engaged in AI development and implementation. From these accounts of AI successes and failures, we propose the need for a broader, infrastructurally-informed understanding not only of the technologies themselves, but also the ways in which they are implemented, built, understood, and acted through as a means of building towards more just and socially positive uses of AI.

2 | BACKGROUND

Popular representations of scientific progress often fail to adequately account for or represent the nuance of that progress. Accounts of AI's successes or failures inform popular understanding of that AI, having significant influences on legislation and regulation, as well as ‘prototype accounts’ of the state of the art for researchers and scholars as they grapple with issues of ethics and needed changes to laws and policy. While some systems are popularly seen to exceed the capacity of humans, with IBM's Watson being perhaps the most visible example (Luxton, 2019), others are seen to reinforce existing bias (Skeem & Lowencamp, 2016), or influence the results of elections and other political action (Metcalf, 2018).

The tools and infrastructures enabling innovation in AI do not respond easily to ethics and values inquiry–often their inner workings are opaque even to the researchers operating them (Kraus, Perer, & Ng, 2016). Considering AI from an infrastructural perspective provides unique leverage into understanding it not only in the form of a specific application, but as the most visible extension of an ‘internetwork’ of data sources, standards, organizations, and applications. While cyberinfrastructure initially concerned itself primarily with computational resources (Atkins et al., 2003), ethnographic research into infrastructure showed that equal attention must be paid to social factors such as group membership, accepted practice, and policy objects like standards or regulation in order to adequately account for its reach and effects (Edwards, 2010; Jackson et al., 2007; Star & Ruhleder, 1996). Infrastructure, in short, comprises that which supports and subtends a given activity—in part it comprises the policy, material and systems that do not need to be reconsidered at the moment of action (Slota & Bowker, 2017). As such, infrastructure becomes a relational quality determined in part by the daily practice and assumptions of availability of a given group.

There have been standards, frameworks and guidelines proposed for ethical AI systems (Bryson & Winfield, 2017; Floridi et al., 2018, Jobin, Ienca, & Vayena, 2019; O'Sullivan et al. 2019; Winfield, 2019). Such standards are often located at the design of algorithms and automated learning techniques rather than addressing contingent systems across the ecosystem of AI, which encompasses data collection, selection, and curation even before concerns of use and guidance in how to interpret and respond to algorithmic outputs (Whittlestone et al., 2019). Professional and technical standardization is used to reduce professional risk, while providing constraints on work practice and homologies between research settings and representations of phenomena (Fujimura, 1992). For example, engineers earn professional engineering licenses and use approved design codes. Standardization allows peers to judge the success and objectivity of work (Sismondo, 2010). Standards concerned with values-oriented societal outcomes, such as those for sustainability or fairness, may also be used in this manner, giving designers and regulators a basis on which to incorporate these values without being concerned about indefensible outcomes or processes that could result in liability concerns. Standards, as they support ongoing work without the need for reconsideration at the moment of that work, are best understood infrastructural goods, part of the inter-network (Slota & Bowker, 2017) that, often in an invisible, occluded way (Slota, Slaughter, & Bowker, In Press), form and structure work and the outputs of technology.

However, standards must also be both learned and interpreted (unlike, for example, the standardized measurement units or physical engineering hardware that are relatively universal [Alder, 1998; O'Connell 1993]). Value Sensitive Design (Fleischmann, 2014; Friedman, 1996; Friedman & Kahn, 2002; van den Hoven, 2007) and Values in Design (Knobel & Bowker, 2011; Nissenbaum, 2001) provide means for understanding and guiding the technology design process according to an understanding of the implicit and inherent values of a system. While these methods are effective in exposing and reasoning through such implicit values, they often present a challenge in terms of selecting and intentionally building systems that conform to or replicate those values in practice (Manders-Huits, 2011), and even in some cases of creating techniques for translating values into design requirements are proposed, it is with the caveat that the relationship is context-dependent, and still requires a values and specification judgement to be made (van de Poel, 2013).

AI is increasingly a distributed, collaborative proposition, and one that, while becoming more consequential, is increasingly opaque in terms of its ethics and values. Standards themselves are designed objects, embedding and reflecting some set of values of their own. While understanding and uncovering embedded values in the design process is a key first step in understanding the technology itself as well as providing some account of its consequence, the infrastructure of its data, standards, policy and legislation also plays a consequential role. The development of standards and values inquiry often focuses solely on the designers themselves. Without some account of the ways in which in the values of systems designers interact, compare and coincide with the values of stakeholders elsewhere in the ecology of AI, efforts to understand and account for the consequence of (or even the full set of embedded values within) AI provide a less-complete picture. Understanding the societal consequence of AI relies upon an account not only the designers of the technology, but at the very least also those who are involved in designing and implementing regulations and laws governing these technologies.

The goal of this paper is to better understand the landscape of the law, regulation, and ethics as it relates to AI through the experiences and work practices of AI stakeholders. Specifically, in this paper, we ask: How can we understand the societal benefits and harms of AI?

3 | METHODS

This paper reports findings from a study in which we conducted 26 semi-structured interviews with an interdisciplinary group of AI stakeholders drawn from the fields of AI research, governmental and organizational policy, and legal research and practice. Here, we define stakeholders as influencing or being affected by AI (Freeman, 1984), which we further partition into technology stakeholders involved in designing AI-based technologies, policy stakeholders involved in the regulation of AI-based technologies, and legal stakeholders involved in interpreting laws in relation to AI. Of these interviews, eight were conducted with technology stakeholders, 10 with policy stakeholders, and eight with legal stakeholders. Potential interviewees were identified according to their engagement with AI and recruited via e-mail.

We took inspiration from real-time technology assessment (Guston & Sarewitz, 2002), using concrete cases of actual technological advances and potential future technological advances that grounded our interviews, rather than assuming legal and policy experts have the relevant technical expertise or are aware of the line between what is possible and what is not. For the latter, we drew inspiration from value sensitive design (Friedman, Kahn, & Borning, 2008), particularly the Envisioning Cards (Friedman & Hendry, 2012), which encourage technology experts to consider the downstream effects and implications of the technologies that they create, and to consider the values of users and others who may be affected by technologies in the design of those technologies. In this way, the interview protocol worked to surface embedded values, elucidate work practice as relevant to negotiating these values, and account for differences in knowledge across the three stakeholder groups. We then used the findings from the technology stakeholder interviews to select specific cases of technological advances to ground interviews with policy and legal stakeholders.

Technology development stakeholders were prompted to consider value tradeoffs (Fleischmann & Wallace, 2010), stakeholder relationships, and the broader potential ethical implications of their own work. Based in part on the findings from these interviews, we identified six emerging applications of AI, including (a) autonomous vehicles, (b) AI-determined organ transplant priority lists, (c) the use of AI agents in call centers, (d) AI-determined decisions on bank loans, (e) the use of AI in medical diagnoses and treatment and (f) the use of AI in informing criminal sentencing decisions. These emerging applications were used in the policy and legal stakeholder interviews, orienting policy and legal interviewees towards technologies with known scope and capacity. For each scenario, we prompted legal and policy stakeholders to consider the ethical, legal, and policy ramifications of that technology. All subjects were also asked to consider the potential consequences of AI in broad adoption in terms of its ethical, legal, and policy implications.

We employed thematic analysis (Braun & Clarke, 2006), first coding the data and then inductively identifying themes across the codes. This analysis resulted in multiple broad themes. This paper reports findings related to one of these themes, the attribution of failures and successes of AI to data quality issues, but we anticipate publishing future papers on AI ethics as work as well as how AI reconfigures the relationship between the individual and the collective.

4 | FINDINGS

Brought to light in these interviews was the role of *accounts* of AI's consequences and implications that bear significant weight in informing how we might better understand, implement, and legislate public accountability in the design and deployment of AI, particularly when our goal is the development of systems that support societally beneficial outcomes. One key account was our stakeholders' depiction of the relationship between the data used to build AI systems and the outcomes of that system. The data, especially, was perceived by many to be substantially out of their control (other than in selection and collection) and a major limit to what their work was capable of doing. One technology stakeholder expressed how what the data expressed was often perceived to be outside the control of the person doing the design work:

> The data is what you got. I mean, [LAUGH], if the data's biased then your result's gonna be biased. If the data's not, it won't, right? So there's some element here beyond, it's beyond the data scientists that guy or gal writing machine learning algorithms, right? It's back to the very data collection itself.

This lack of control at the moment of analysis and design of AI systems is further emphasized by our interviewees attribution of negative social consequences such as bias, unfair liability, or reinforcing of unwanted social conditions as largely unavoidable due to the quality and coverage of available data. Throughout our interviews with stakeholders in AI, there was a persistent notion that negative social outcomes of AI systems can be attributed, not to the analysis, algorithms or system design, but to the data itself. In attempting to assess or address societal outcomes of algorithmic analysis, data quality and availability was seen as a significant limiting factor in what designers and researchers in AI systems are able to do and one that bore substantial consequence as it made its way into the world. An AI expert stated this quite succinctly:

> As you can imagine, that model is limited to the data that it processed, the data it was trained on. That's its entire universe of knowledge. So if that data set is incomplete or if it's biased, it's going to result in a biased result for the end user.

As seen above, both the quantity of data necessary, and its often siloed, heterogeneous nature was seen as a significant hurdle in implementing ML in particular. Some interviewees, such as the technology stakeholder below, emphasized the need for broader, interoperable databases in order to achieve desired results in their commercial AI environments.

> And so to accomplish that we need lots and lots of data. That's one of the big issues of machine learning, particularly deep learning, lots of data. And so we have all these internal silos, different business units that have databases.

However, complexity in existing analysis, specifically when adapting AI towards socially positive outcomes, can also be limited by existing practices. This becomes not just an issue of data and data quality, but also of engaging with domains, particularly in their histories of ongoing work. The climate model, in particular, was described by a technology stakeholder engaged in AI research as being particularly intractable to machine learning approaches due not only to the data itself but also the history of work undertaken within that domain:

> The problem is that most of the data sets being collected are very small, fragmented, not necessarily the type that are easy to learn from. Our climate models have become very, very, very complicated over the years and replicating them with machine learning techniques is actually really difficult.

This notion significantly problematizes standardization and regulation efforts that focus on the moment of algorithmic design itself. Undesirable and biased social conditions were understood by our interviewees to affect the data that is available, and this was seen, as expressed below by a technology stakeholder engaged in commercial research, as making AI itself more applicable and effective for already-privileged groups:

> And I think that the sort of overwhelming message from deployed machine learning over the past five years is that it can still work much, much better. For people who are in socially advantaged classes like men. People who are economically well-off, people with lighter skin and so on, and tend to work comparatively poorly for groups who are either under-represented or disadvantaged.

However, these biases in data are often invisible until after analysis takes place. It is difficult to identify or account for biased data prior to analysis, and often such incomplete data coverage remains even after the technology is deployed. This was echoed by a legal stakeholder in AI, where bias became notable as 'in-built' on the basis of available data:

> There is latent bias, you see that a lot ... Where the data set that the model is trained on simply didn't include enough people of particular, for example, racial heritages and so it doesn't work very well because they didn't have enough. And then you can also see bias built in.

And the fact remains that implementations of AI outside of the research space often retain those biases. Researchers and designers might consider them irrelevant to that application, or impossible to correct. The distance from control over outputs as expressed above becomes increasingly problematic as innovation, novelty and being the first to market become prioritized (especially in commercial contexts) over working with unbiased, well-covered data sets. Despite the problematization of data availability, respondents also expressed the notion that biased, incomplete data might not necessarily negatively impact the results of research or technology design. One technology stakeholder in a leadership role of a technology company referred to some difficulty in assessing how impactful biased data might be in certain applications - or even whether that bias might be desirable.

> How do we go about making sure our data isn't biased in some way? Or even deciding, maybe we've figured out that it's biased but we decide that that's okay. In our application, that doesn't matter. Maybe it matters very much.

This stakeholder added:

> it's possible, because we're selling computers, that we get our data off our own website and only technical geeks [LAUGH] like to buy computers. So the very data that we collect is biased. Not because we're evil people, but because that's just what happens in the real world, right? Certain people like to buy our stuff.

However, even among those interviewees who were aware of algorithmic accountability, and sensitive to standards for ethical policy design, adequately representing bias was itself problematic, as expressed by the following technology stakeholder.

> Right now maybe some of the bad effects of AI that we see, things like biased AI algorithms that demonstrate sort of racist or genderist sort of biases. I guess maybe it's just an outcome - we don't have the tools to express that.

And the representation of that idea was perceived as quite important, especially when discussions of regulation and the legal implications of the outcomes of AI occurred. While those working directly in the design process were aware of the limits both of their algorithms and the data that they used, they lack control of how those results are perceived, and even whether those limitations would be communicated or discussed once the designed system made its way to an end-user or the results became publicized. One interviewee related their experience in writing a paper intended for the security community that expressed some shared narratives of unexpected outcomes of algorithmic analysis and itself gained media attention and found traction outside of the initially intended community. Upon collecting and publishing these stories, with a goal of representing that knowledge that "*didn't fit the narrative*," of the scientific process, this interviewee found their paper gaining significant traction outside of the original intended community, including public media.

> And I guess kind of, because the paper's been amplified in the media, it's had some diffuse effect there. And the media sometimes distorts things and so it could be unclear whether the message has gotten through in the way that we intended.

While this interviewee was generally positive about the reception and potential influence of the paper, they were still concerned about distortion and misrepresentation once it moved to other contexts, was drawn upon by other communities. Hype, as it emerges as a theme in our analysis, thus expresses not just an ill-informed, future-oriented representation of the potential for a technology or the results of research, nor a breathless prediction of the future, but in general the notion that outcomes of AI are often distorted as they leave the immediate community for which it they were intended. This interviewee was made very aware of public perception of their research and design work, and became quite concerned with how it was represented.

> Things that they said that were very precise statements kind of got taken out of context … I assumed that people would always read the full paper and get the full context. They want to cherry pick certain lines, and that's what happens in the real world, so I wasn't super prepared for that.

The "messiness" of incompatible understandings of the capacity of AI between those working with it more directly and those receiving information outside of the context of that direct work was seen by many interviewees, including the technology stakeholder below.

> There's also the negative impact of, because different media end users are more or less going to, are going to more or less distort the message as some academic thing, more or less. There's some false ideas that are going out as a result as well, so kind of messy.

This quote presents an interesting contrast to the accounts presented by the designers of technology themselves of the successes and failures of their systems. While they understood biased results arising from their own work as resulting from nuanced issues of data coverage and the presence of unexpected outcomes in algorithmic design, their perception of the publicity of those results was often oriented towards media figures and popular press misunderstanding. In particular it was media hype of the type mentioned in the above quote that was seen as misleading. And this distortion created an interesting parallel to the expressed incompleteness or lack of coverage of the data itself. Much as outcomes of AI were seen as escaping the control of the designers due to incomplete, biased, or inaccurate data, the perceptions of those outcomes were similarly perceived as outside of the control of designers and researchers. A technology stakeholder working in the field of autonomous vehicles explicitly called out this lack of control as needing some form of address.

> I guess, one place I'm curious about, I'm not sure what could be done better or worse, is the interface between research and the media. There's a lack of control there. I'm not sure what sort of intervention could be done.

This interface between media and research, especially in conveying and representing an accurate understanding of these results was perceived as consequential. A legal stakeholder drew a connection between the representation the media hype of AI (in a negative sense as 'fear-mongering') and legal consequences:

> If you look at first of all the fear-mongering that's going on regarding Artificial Intelligence and all of the disinformation out there, and you look also at the few laws that were adopted, it seems that the burden is being put more on the AI users, and when I say user, I mean the companies that will actually

> buy AI solutions and incorporate them into their products.

The above quote raises a particularly salient issue: technology and scientific experts, while they have a role in law and policy formation (Jasanoff, 2004) are not the only means by which the public and policymakers understand technology. Negative media coverage and incomplete or inaccurate representations of results or process might create unrealistic expectations on the part of the public (or even students entering the field), and contribute to placing the burden of accountability on users that have little control of issues like biased or incomplete data, and fail to sufficiently account for the process by which outcomes of AI come to be. A technology stakeholder mentioned the negative consequence such media hype might have on the field of machine learning itself:

> My most negative concern is that people have over hyped what machine learning and what artificial intelligence can do. I've read a lot of pop science articles on, The development of, what do they call it? Artificial general intelligence. And I've seen a lot of grandiose claims. And I think a lot of those projects are kinda bound to fail. And if they succeed that's great. It would have a huge impact, but they might fail. In terms of the machine learning space, I think when people start pointing to these failures, they're gonna grow cold on the things that machine learning can actually achieve which again, I think that's a lot.

In this, we see another factor impacting the social consequence of AI—that of its popular understanding, and the generally available knowledge of what its capabilities and capacity might be. Hype pushes the discourse of ML/AI towards unrealistic questions, and popular representations in the media might relocate responsibility to those with less control of the outcomes. We perceive this as a problem of the *socialization* of AI—here conceived as the availability and pervasiveness of an infrastructure by which AI might be publicly understood, evaluated, and held accountable. A well-socialized technology is one that is consistently represented, where sufficient implicit knowledge exists that misrepresentation is relatively visible, and misunderstandings can be more readily corrected. More specifically, the socialization of technology speaks to its visibility and in shared understandings of its role. And the lack of socialization can be problematic, with interpretability being of central concern to the users of that technology, as the below legal stakeholder states with respect to issues of liability as learning-based expert systems find their way into high-risk applications like medicine.

> I think that you will see claims that the AI was mistaken, so any time a person is injured through a medical procedure, they have the possibility of asserting malpractice. And you start to raise questions of was the clinician competent to rely on the AI, AI can be a complex technology to understand. And if you do not understand how the model was trained and what the target of the model is, it can easily be misunderstood and misapplied potentially resulting in human injuries and of course losses.

As seen in the following quote from a legal stakeholder, how AI is framed and understood popularly can not only provide unrealistic views of the capacity of AI, but also limit and structure where they might understand ethical impacts to occur.

> Unfortunately, for most people, AI is framed for them by Hollywood. And they think Terminator, and various movies, and they think general artificial intelligence. And that drives a lot of the thinking and conversation around ethics. No one would say, hey, we've really got to have an ethical program in Microsoft Word, it's a piece of software, there's no sentience. The reality of AI today is, there is no sentience … So we really need to understand what we're talking about, as we start to think about things like education. And it's sort of, everyone needs to understand what the technology is, in order to develop a rational educational framework.

The stakeholder above, in talking about how a "rational educational framework" might be put into place, clearly elucidates the need for better socialization of AI even prior to educational interventions. When 'sentience' or some other criteria becomes a dividing factor between where values and ethics ought and ought not to be considered in system, the implicit values and societal impacts of those technologies that fall on the 'ought not' side of that criteria are occluded, and potential mitigations of negative impacts may remain unconsidered. Algorithmic operation cannot be fully understood as separate from the data on which it operates, neither can data be considered as separate or distinct from its representation, analysis, and deployment. Apparent neutrality of technology is produced through limited representation—all technology is sociality as well as design, and an effective understanding of the societal impact or consequence of technology requires a similar understanding of how it is socialized, regulated, and used.

5 | DISCUSSION

What do our findings tell us about how to understand the societal benefits and harms of AI, specifically in terms of maximizing benefits and minimizing harms? There is no such thing as truly objective or neutral design—even very mundane objects can be shown to embody some set of values and play vital social roles (Latour, 1992; Winner, 1980). The lack of control over outcomes expressed by our interview subjects indicates that it is not just data, or algorithms, or implementation, rather a connected, complex view of their interactions are necessary for a better understanding of the social ethical consequence of a system. While data was often pointed to as a key factor in biased or otherwise flawed applications of AI, this evidence also indicated that better data will not necessarily result in a more ethically sound system. Public interpretations of the capacity and action of the system will play in a role in how it is regulated, understood and funded. Systems with 'perfect data' for their intended purpose might still be misused or misunderstood, and their potentially positive effects undermined.

Our interviews revealed a very close conceptual relationship between the outcomes of AI and the specific data sources used to inform its development. Much as early expert systems and other applications of symbolic AI required a close engagement with the domains and experts that they are intended to serve (Ribes et al., 2019), our technology stakeholders engaged with modern, learning-oriented and algorithmically-driven AI considered data not only as a resource but also as a key determining factor in the effectiveness and outcomes of AI both in research and in implementation outside of research contexts.

In our interviewees' common attributions of negative social outcomes of algorithmic analysis to problematic data, it might seem that better data would be the solution to producing better, more ethically sound systems. However, it is often difficult to control the outcomes of technology design and implementation even when attention is paid to the embedded values, ethical standards and regulation of AI. Very visible outcomes, like IBM's Watson's run on Jeopardy, or the ability to use social data (often illicitly obtained) to broadly influence elections based on models of human behavior (Metcalf, 2018) often fail to represent or account for the process and operation by which those results were obtained. These influential, incomplete representations and implementations of AI exceed the control and contextualization of their designers. We have chosen to refer to this as hype in order to both make use of the members' meanings (Emerson, Fretz, and Shaw, 1995) of the term and emphasize the expressed distance between the designers' understandings of their technology and the ways in they saw that technology being perceived and expressed in the larger public sphere.

The findings presented in this paper point towards the need for an ecological, infrastructural understanding of AI as being necessary to effectively assess or mitigate its potential social effects. A question often raised in discussions of transparency or accountability of algorithms is 'transparency for whom?, accountable to whom?' (Ananny, 2016; Kemper & Kolkman, 2019; Kroll et al., 2016). Moreover, what is needed is not just an infrastructural view of the ethical and societal ramifications of AI, but also infrastructures of social good. Building socially good AI was often cast in our interviews as somewhat outside of the purview of the technology stakeholders involved in its development, and similarly a lack of full understanding of a given "mode" or implementation of AI was pointed to by policy and legal stakeholders as (mis)apportioning responsibility for the outcomes of that technology or being significant limiting factors in potential societal good that they might produce. Occluded, marginalized, or embedded aspects of policy, technology, and society bear consequence in terms of which human values the system supports, enacts, or challenges—even in what might seem to be relatively ancillary aspects of that technology like terms of service agreements (Slota, Slaughter, & Bowker, In Press). As such, building an ethically good AI system is an exercise in building and exploring infrastructure, not only to better understand how a negative or positive societal outcome might come to be in the confluence of data, analysis, implementation and regulation, but also in providing a means of understanding AI accurately in terms of its scope, capacity, and potential limitations.

The socialization of AI is not limited to the educational framework and availability of knowledge, but rather encompasses the broader set of systems, policies, and standards that render the social outcomes of AI available to regulation, tractable to policymakers and the public at large, and the shared vocabularies and understandings that can limit the negative effects of incomplete representations of that data. An example of a well-socialized technology might be that of the automobile. While the design and manufacturing process of a given vehicle might not be immediately available knowledge to a large group of people, the responsibilities of drivers and manufacturers are relatively well-understood and accepted, and infrastructure exists to maintain, ensure the safety of, and support their operation. Necessary infrastructural goods like fuel are present, standards (though themselves living and changing) are present and enforced, and the broad strokes of the required knowledge and competencies to make safe and effective use of the technology are established and fairly consistent. While some might claim widespread driving is itself a social negative, that judgment call varies widely—more importantly to this is the notion that should a person choose to drive or build a car there is an accessible, supported, and maintained means by which they might understand their responsibilities, liabilities, and requirements for safe operation.

A prescriptive ethics of AI may not be the immediate goal in seeking social good and ethically responsible uses of that technology. Rather, a means by which an effective, informed discussion on what is ethical or unethical in the uses of AI is a necessary prior step, and one that (especially given the presence of hype and its potential impacts on policy and regulation) cannot take place solely in the academy.

6 | LIMITATIONS AND FUTURE DIRECTIONS

This exploratory study helped us to identify both findings and limitations. First, our interview group was largely self-selected, and likely was self-selected towards those stakeholders who already had some interest in the ethics or values of AI systems, given the time requirement and lack of compensation. Second, as an interview-driven study, we did not directly examine code, nor did we deeply explore research or work practices *in situ*, and largely took our interviewees accounts of their work at face value. Third, while our approach was effective in eliciting responses informed by an effective understanding of the capacity of our chosen AI technologies for our legal and policy stakeholders, as well as sensitizing our technology stakeholders to issues of values, ethics, law, and policy, this approach had limited impact beyond our immediate studied population.

Future research could potentially engage with a wider population of stakeholders through more lightweight methods. For example, surveys could be distributed via professional societies to gain an understanding of how these findings could apply to the broader community of legal, policy and technology stakeholders of AI. Our study design also revealed significant opportunities for research that more directly engages with work and research practice through participant ethnography, real-time technology assessment (Guston & Sarewitz, 2002), or participatory design activities. Finally, the techniques we developed to coordinate understanding between our stakeholder groups could inform the development of educational and training interventions that could be adopted to do so on a larger scale.

7 | CONCLUSION

We started with the question, how can we understand the societal benefits and harms of AI? Our participants revealed a broad picture of responsibility and agency in the impact of AI, that extended beyond the research into and design of AI, its regulation, and relevant legal structures. Understanding how some outcome came to be requires an understanding of the full lifecycle of the technology that gave rise to the outcome, from data collection, curation and selection, all the way to how that system comes to be represented and understood in the media. In similar ways, policy responses to and the regulation of AI likely will be significantly limited in effectiveness without a full accounting of the infrastructure that subtends the work of AI. While there exist a wide array of current efforts to standardize, provide design direction, and coordinate research and implementation work towards more socially positive outcomes of AI, these often tend to place responsibility and accountability for those outcomes at the moment of design, or at the end user of the system.

Data-driven AI, especially when heterogeneous data sets are collated and leveraged towards learning applications, significantly complexifies the agency of the designer or user of a given AI system in terms of its potential outcomes. In addition to the quality, quantity, and coverage of data, and the media representations of the capacity and potential of a given application of AI, there still exists a landscape of regulation, extant standards, professional expectations, and design affordances that all contribute to how AI is used and what its impact might be. In order to build good systems, there needs to exist an infrastructure through which the impact and ethical valence of the ecosystem of AI might be effectively understood.

ACKNOWLEDGMENTS
This research was funded by Cisco Systems under RFP-16-02, Legal Implications for IoT, Machine Learning, and Artificial Intelligence. We also thank our research participants in the larger study, including Blake Anderson, Ruzena Bajcsy, Hal Daume III, Stephen Elkins, Enzo Fenoglio, Iria Giuffrida, Dean Harvey, James Hodson, Wei San Hui, Amir Husain, Jeff Kirk, Frederic Lederer,

Ted Lehr, Terrell McSweeny, Matt Scherer, Peter Stone, Nicolas Vermeys, and Christopher Yoo as well as eight participants who elected to remain anonymous.

REFERENCES

Alder, K. (1998). Making things the same: Representation, tolerance and the end of the Ancien Régime in France. *Social Studies of Science*, *28*(4), 499–545.

Ananny, M. (2016). Toward an ethics of algorithms: Convening, observation, probability, and timeliness. *Science, Technology, & Human Values*, *41*(1), 93–117.

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messerschmitt, D. G., … Wright, M. H. (2003). *Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*. Washington, DC: National Science Foundation. Retrieved from https://www.nsf.gov/cise/sci/reports/atkins.pdf

Barocas, S., & Boyd, D. (2017). Engaging the ethics of data science in practice. *Communications of the ACM*, *60*(11), 23–25.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101.

Brogaard, J., Hendershott, T., & Riordan, R. (2017). High frequency trading and the 2008 short-sale ban. *Journal of Financial Economics*, *124*(1), 22–42.

Bryson, J., & Winfield, A. (2017). Standardizing ethical design for artificial intelligence and autonomous systems. *Computer*, *50* (5), 116–119.

Cartea, A., Donnelly, R., & Jaimungal, S. (2017). Algorithmic trading with model uncertainty. *SIAM Journal on Financial Mathematics*, *8*(1), 635–671.

Chaboud, A. P., Chiquoine, B., Hjalmarsson, E., & Vega, C. (2014). Rise of the machines: Algorithmic trading in the foreign exchange market. *The Journal of Finance*, *69*(5), 2045–2084.

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. Cambridge, MA: MIT Press.

Emerson, R. M., Fretz, R. I., & Shaw, L. L. (1995). *Writing ethnographic fieldnotes*. Chicago, IL: University of Chicago Press.

Fleischmann, K. R. (2014). *Information and human values*. San Rafael, CA: Morgan & Claypool.

Fleischmann, K. R., & Wallace, W. A. (2005). A covenant with transparency: Opening the black box of models. *Communications of the ACM*, *48*(5), 93–97.

Fleischmann, K. R., & Wallace, W. A. (2009). Ensuring transparency in computational modeling. *Communications of the ACM*, *52*(3), 131–134.

Fleischmann, K. R., & Wallace, W. (2010). Value conflicts in computational modeling. *Computer*, *43*(7), 57–63.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707.

Freeman, R. E. (1984). *Stakeholder management: Framework and philosophy*. Mansfield, MA: Pitman.

Friedman, B. (1996). Value-sensitive design. *Interactions*, *3*(6), 16–23.

Friedman, B., & Hendry, D. (2012). *The envisioning cards: A toolkit for catalyzing humanistic and technical imaginations*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1145-1148).

Friedman, B., & Kahn, P. H., Jr. (2002). Human values, ethics, and design. In J. A. Jacko & A. Sears (Eds.), *The human-computer interaction handbook* (pp. 1209–1233). Mahwah, NJ: Lawrence Erlbaum.

Friedman, B., Kahn, P. H., & Borning, A. (2008). Value sensitive design and information systems. In P. Zhang & D. Galletta (Eds.), *The handbook of information and computer ethics* (pp. 69–101). Armonk, NY: M. E. Sharpe.

Fujimura, J. H. (1992). Crafting science: Standardized packages, boundary objects, and "translation.". *Science as Practice and Culture*, *168*, 168–169.

Guston, D. H., & Sarewitz, D. (2002). Real-time technology assessment. *Technology in Society*, *24*(1-2), 93–109.

Jackson, S. J., Edwards, P. N., Bowker, G. C., & Knobel, C. P. (2007). Understanding infrastructure: History, heuristics and cyberinfrastructure policy. *First Monday*, *12*(6).

Jasanoff, S. (Ed.). (2004). *States of knowledge: The co-production of science and the social order*. New York, NY: Routledge.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399.

Kemper, J., & Kolkman, D. (2019). Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society*, *22*(14), 2081–2096.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 2053951714528481.

Knobel, C., & Bowker, G. C. (2011). Values in design. *Communications of the ACM*, *54*(7), 26–28.

Kranzberg, M. (1986). Technology and history: "Kranzberg's Laws." *Technology and Culture*, *27*(3), 544–560.

Krause, J., Perer, A., & Ng, K. (2016). *Interacting with predictions: Visual inspection of black-box machine learning models*. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 5686–5697).

Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2016). Accountable algorithms. *University of Pennsylvania Law Review*, *165*, 633.

Latour, B. (1992). Where are the missing masses? The sociology of a few mundane artefacts. In W. Bijker & J. Law (Eds.), *Shaping technology, building society* (pp. 225–258). Cambridge, MA: MIT Press.

Lehr, D., & Ohm, P. (2017). Playing with the data: What legal scholars should learn about machine learning. *UCDL Review*, *51*, 653.

Luxton, D. D. (2019). Should Watson be consulted for a second opinion? *AMA Journal of Ethics*, *21*(2), 131–137.

Manders-Huits, N. (2011). What values in design? The challenge of incorporating moral values into design. *Science and Engineering Ethics*, *17*(2), 271–287.

Metcalf, J. (2018 April). Facebook may stop the data leaks, but it's too late: Cambridge Analytica's models live on. *MIT Technology Review*. Retrieved from https://www.technologyreview.com/s/610801/facebook-may-stop-the-data-leaks-but-its-too-late-cambridge-analyticas-models-live-on/

Nissenbaum, H. (2001). How computer systems embody values. *Computer*, *34*(3), 120–119.

O'Connell, J. (1993). Metrology: The creation of universality by the circulation of particulars. *Social Studies of Science*, *23*(1), 129–173.

O'Sullivan, S., Nevejans, N., Allen, C., Blyth, A., Leonard, S., Pagallo, U., … Ashrafian, H. (2019). Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. *The International Journal of Medical Robotics and Computer Assisted Surgery*, *15* (1), e1968.

Ribes, D., Hoffman, A. S., Slota, S. C., & Bowker, G. C. (2019). The logic of domains. *Social Studies of Science*, *49*(3), 281–309.

Sismondo, S. (2010). *An introduction to science and technology studies* (Vol. 1). Chichester, England: Wiley-Blackwell.

Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, *54*(4), 680–712.

Slota, S. C., & Bowker, G. C. (2017). How infrastructures matter. In C. Miller, L. Smith-Doerr, & U. Felt (Eds.), *The handbook of science and technology studies* (pp. 529–554). Cambridge, MA: MIT Press.

Slota, S. C., Slaughter, A., & Bowker, G. C. (In Press 2020). The Hearth of Darkness: Living Within Occult Infrastructures." Forthcoming in L. Lievrouw & B. Loader (Eds.), *Handbook of digital media and communication*. London, England: Routledge.

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, *7*(1), 111–134.

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J. … Teller, A. (2016). *Artificial Intelligence and Life in 2030*. One hundred year study on artificial intelligence: Report of the 2015-2016 Study Panel. Stanford University, Stanford, CA. Retrieved from http://ai100.stanford.edu/2016-report

van de Poel, I. (2013). Translating values into design requirements. In I. van de Poel & D. E. Goldberg (Eds.), *Philosophy and engineering: Reflections on practice, principles and process* (pp. 253–266). Dordrecht, Germany: Springer.

van den Hoven, J. (2007). ICT and value sensitive design. In P. Goujon, S. Lavelle, P. Duquenoy, K. Kimppa, & V. Laurent (Eds.), *The information society: Innovation, legitimacy, ethics and democracy in honor of Professor Jacques Berleur SJ* (pp. 67–72). Boston, MA: Springer.

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). *The role and limits of principles in AI ethics: towards a focus on tensions*. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 195-200).

Winfield, A. (2019). Ethical standards in robotics and AI. *Nature Electronics*, *2*(2), 46–48.

Winner, L. (1980). Do artifacts have politics? *Daedalus*, *109*, 121–136.

How to cite this article: Slota SC, Fleischmann KR, Greenberg S, et al. Good systems, bad data?: Interpretations of AI hype and failures. *Proc Assoc Inf Sci Technol*. 2020;57:e275. https://doi.org/10.1002/pra2.275