Trustworthy Algorithmic Decision-Making*

Workshop Report

Emilee Rader and Rick Wash

December 4–5, 2017

Executive Summary

Computer-based algorithms are increasingly being used in systems that automatically make important decisions on behalf of people, including determining what news people see online, controlling speed and steering of cars, choosing prices for goods and services, filtering job applicants, recognizing and categorizing airport travelers, and making sentencing recommendations for people convicted of crimes. As these algorithms simultaneously become more common and more complicated, it is important to understand whether they can be trusted to make decisions like these, what makes algorithms trustworthy, and how algorithms can be made more trustworthy.

Fundamentally, these algorithms operate in a complicated socio-technical context that includes the designers of the algorithms, the data used as an input to the algorithms, the interface that presents and uses the outputs, the people who make choices about goals of algorithms and when to use algorithms, and societal laws and norms that influence their use. All aspects of this context influence the outputs of the algorithms, and also impact whether they are worthy of being trusted to make important decisions.

A group of researchers, practitioners, and policy-makers convened at a workshop on December 4–5, 2017 in Arlington, VA to discuss these issues. As a group, we identified five challenges that future research needs to focus on to help algorithms be more trustworthy:

1. *People, processes, and training*: Who defines and how do we ensure good practice in data science and machine learning? What are the avenues for education? What are the appropriate tools for ensuring good practice in machine learning and algorithm development?

2. *Evidence, accountability, and oversight*: How do we integratively assess the impact of an algorithmic system on the public good?

3. *Handling uncertainty*: How do we holistically treat and attribute uncertainty throughout data analysis and decision making?

4. *Adversaries, workarounds, and feedback loops*: How should trustworthy algorithms account for and be resilient to adversaries who try to manipulate the algorithms, workarounds from people trying to achieve their goals, and feedback loops where algorithm outputs become future inputs?

5. *How do we trust algorithms*? What are the processes through which different stakeholders come to trust an algorithm?

*Funded by the National Science Foundation under grant No. 1748381. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Contents

1. Algorithms and Decision Making

Systems that use algorithms and large datasets to automatically make decisions on the behalf of users are becoming more integrated into everyday life. For example, they do things like determine what news people see in social media [9] and even generate some news reports without human intervention [11]; direct cars to react to changing conditions around them via autopilot features [12]; govern differential pricing of goods and services [14]; speed up the airport check-in process using facial recognition [4]; filter job applicants [3]; and make risk assessments and recommendations for sentencing people convicted of crimes [13]. The key feature that all of these examples have in common is that decisions which used to require human judgment and agency are now being made by what is in many cases a "black box": a "system whose workings are mysterious; we can observe its inputs and outputs, but we cannot tell how one becomes the other" [10].

The word "algorithm" used to be specialized jargon in mathematics and computer science, and referred to any process that receives input, operates on the input, and produces an output [6]. In recent years, we have seen the meaning and usage of this word transform and emerge into popular discussions of new technologies. It is increasingly being used colloquially to refer to systems like those in the preceding examples, which are becoming more complex and require statistical and machine learning models that find patterns in large datasets and make decisions based on those patterns. Computing systems are approaching Licklider's vision of man-computer symbiosis [8]: offloading tasks that computers are good at and freeing humans to focus on problems computers are unable to solve. As people come to rely more and more on systems like these, it is important to ensure that they are *trustworthy*: that they perform in known, predictable ways and are able to be held to legal, social, ethical, and technical standards of performance.

A trustworthy system is one that people can rely on to perform as expected [7], even in uncertain situations when it would be difficult for a human to choose how to proceed. Trust is a key factor that helps people decide whether to use systems that engage in algorithmic decision-making [5]. However, it can be difficult for people to determine whether a system is trustworthy, because in many cases it is not possible for end users to interrogate these systems and figure out how they work. Many of the details are carefully guarded corporate secrets, but also, sometimes even the designers and engineers responsible for creating these systems can't predict the solutions or decisions they produce [1]. In addition, users are often unaware of all of the various stakeholders and partnerships that go into provisioning the data the algorithms use, are unable to find out what kinds of outputs the system is optimized to produce and therefore cannot assess whether it is working the way it is supposed to, and since the outputs are often personalized or specifically tailored to a given situation comparing one's experiences with other users does not yield useful insights.

Ensuring that systems that incorporate decision-making algorithms are trustworthy is also particularly challenging in the face of adversaries: those who intentionally try to affect the inputs—and therefore the decisions—of these algorithms. An old example of adversaries in algorithmic decision-making is search engine optimization (SEO); an entire industry has developed to try to artificially influence the algorithms that filter and rank results from Internet search engines by understanding the algorithms used and altering

or influencing the data (webpages) that is input into the algorithm. What is the equivalent of SEO for self-driving cars? Is it possible for people to alter the data, for example, in order to force cars to alter their routes and drive past specific businesses or advertisements? End users have begun to do just this with Google's Waze app, reporting fake accidents to prevent the routing algorithms from directing commuters through residential neighborhoods [2].

These questions are becoming more critical as three trends are converging: 1) algorithms are increasingly being used to make important decisions that affect people's lives in many ways; 2) algorithms themselves are becoming more complex and outputs more difficult to directly understand and predict; and 3) algorithms are increasingly relying on datasets which have their own biases, are constantly changing, and provide a vector of influence for adversaries.

2. Workshop

In December, 2017 we brought together 42 researchers, practitioners, and policy-makers to discuss "trustworthy algorithmic decision-making". The explicitly stated goal was to characterize the problem space around trusting algorithms in the context of decision-making: what are the major challenges that scholarly, scientific research will need to address to help society understand algorithms and algorithmic decision-making, to create more trustworthy algorithms, and to use algorithms in a more trustworthy way.

Attendees were recruited via a two-stage process. A call-for-whitepapers was distributed through a wide variety of academic and non-academic channels asking for a two-page paper discussing current issues, approaches, or case studies around problems related to trustworthiness in algorithms and algorithmic decision-making. We received 50 whitepaper submissions. We prioritized diversity of viewpoints about these issues, and invited 35 whitepaper authors to attend the workshop. Whitepapers from invited attendees are publicly available on the workshop website: `http://trustworthy-algorithms.org/whitepapers/`.

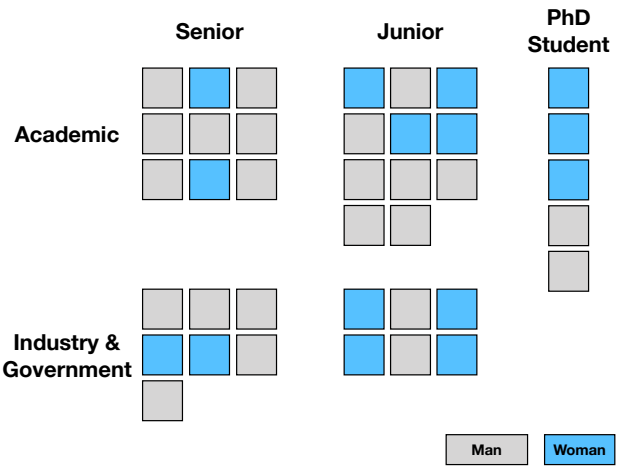


Illustration of the diversity of participants in the workshop. Participants represented a wide variety of disciplines, including computer science, business, economics, social work, communications, information science, statistics, political science, electrical engineering, and geography.

In a second stage, we also extended invitations to additional individuals who did not submit whitepapers but had valuable perspectives on these issues. Many of these invitations went to people who could not submit a whitepaper due to employer contraints (for example, who work for specific industry or government organizations that would require a lengthy approval process). This second stage helped get additional perspectives from industry and government. We ended up with a very diverse group of individuals (see the illustration above) that brought an even more diverse set of perspectives and approaches to the problems.

We convened in the Ritz-Carlton Hotel in Arlington, Virginia on December 4–5, 2017. The workshop began on the first day with some attendees giving short talks about their ideas, with group discussion and note-taking. Over lunch, an affinity diagram was created by the group, and five major themes emerged from the notes. We then separated into breakout groups, with each group discussing and working on clarifying one of the five themes. Attendees rotated through different groups, giving them opportunities to provide input and thoughts on (almost) all of the different themes. On the second day, after more work in breakouts, each group presented to the whole workshop a statement of the problem the group discusses, why it was difficult, why it was important to address, why progress is possible, and potential barriers to success.

Notes were taken through the whole workshop, and were assembled and transcribed after the workshop. This report summarizes these five themes and the challenges identified by attendees around these themes.

3. Challenges for Trustworthy Algorithms

Attendees at the workshop identified and developed five major challenges. All of these challenges are important for making algorithms and algorithmic decision-making more trustworthy. They are also challenges that need new research and ideas to help solve them, but attendees at the workshop believe that this research is possible with appropriate focus, funding, and effort.

3.1. Processes, People, and Training

In his talk during the opening sessions of the workshop, Shion Guha described an instance of people using algorithms for crime analysis. He described a couple of young employees who recently graduated with a masters degree in Criminology (which provides little training in algorithms). They were employed by the local police department, which had recently purchased software that uses a machine learning algorithm (k-means clustering) to help the police predict where crime might happen to allocate police resources better. These two recent graduates ran the clustering algorithm with 5 clusters. Why 5? Because "that's what our sergeant told us."

Participants discussing "people, processes, and training" during the final breakout session.

Algorithms are not used in a vacuum; people use them to accomplish goals. One of the biggest challenges in helping make algorithms more trustworthy is ensuring that the people who are using algorithms have appropriate training and education to make appropriate choices. We cannot rely on highly specialized PhDs and high levels of technical expertise among the people who are making critical decisions about algorithms. It is still important for everyone who works with algorithms to have a basic literacy about bias, fairness, and consequences of their decisions. *Who defines what is best practices around algorithms? And how can we ensure that people follow these practices?*

In order for algorithms and the use of algorithms to be more trustworthy, it is important that the people who are using the algorithms understand not just the algorithm, but also the social, cultural, and societal context that the algorithm is being used in. Decisions about the use of algorithms should not just focus on the immediate goals, but should also take into consideration the effects on end users, on fairness, and on social good.

It would help if the community could come up with a set of best practices that people who are involved in the day-to-day use of algorithms could follow that will help ensure that algorithm use can be trusted by society. Some example best practices may be:

- You should be talking with people who don't look like you, who don't speak like you, and who may be affected by what you do.
- You should be connected with the consequences – practical and social – of your work.
- You should be validating your analyses and models
- You should discuss and expressly communicate the limitations of your work
- You should be outlining constraints and caveats that are built into your work
- You should prefer transparency when possible and valuable

There is an important challenge in defining who these best practices should apply to. What does it mean for a person to "use an algorithm" or be engaged in "data science"? Identifying the set of people who have influence over an algorithm and its uses is difficult, but should at least include the people who choose whether to use an algorithm, the people who choose which algorithm to use, the people who choose which data to use, the people who choose algorithm parameters, the people who validate the algorithm's output, and the people who use algorithmic outputs to make decisions. At a high level, a reasonable rubric is that if you are touching data or algorithms in any way, then these best practices should apply.

Why is this important? This is important because algorithms are being used in ways that can seriously affect people, but are increasingly being used by more diverse groups of people with a wider variety of backgrounds and training. As it gets easier for algorithms to be used by people with less training, it is important to maintain some best practices so that the algorithms are used in a trustworthy way that benefits society.

Why is this difficult? Establishing best practices is difficult because there are many incentives that make it hard to follow these practices. Often there are pressures – time pressure, financial pressure, data limits – that make it difficult to consider alternatives or to really evaluate the impacts of algorithms. Many people who use algorithms work in disparate fields or domains and may not have effective ways of communicating best practices, may have incentives that prevent or limit sharing of best practices, and may use different terminology, jargon, and concepts that make communication difficult. Also, best practices may differ across domains for good reasons, limiting the ability to identify what actually is "best".

Recommendations for Progress One of the biggest recommendations is to find ways for practitioners to share experiences and practices with each other, especially across domains. For example, we could study and identify incentives to reward sharing of data and code examples, or establish a set of case studies that describe mistakes and problems that have occurred.

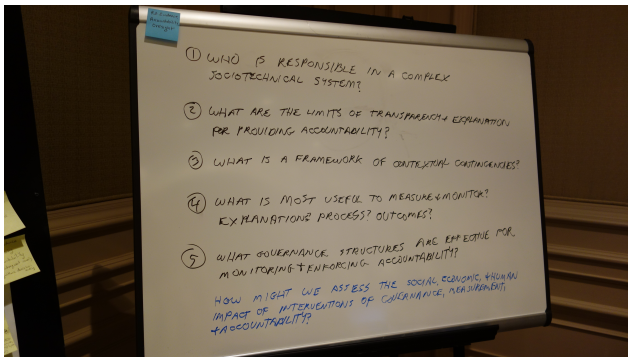
Communication is one of the biggest challenges in training. It would really help for some community (the AI community?) to establish better common vocabularies around algorithms and the impacts that algorithms can have. For example, research that identifies common problems that can lead algorithms to not be trusted would be helpful in creating this common vocabulary and best practices.

It is important that education around algorithms and best practices continue to happen not only within institutions of higher education, but also beyond them. MOOCS, conferences, meetups, professional organizations, and continuing education are all excellent opportunities for algorithm practitioners to share best practices and stay educated about fairness, bias, and trust issues.

Finally, it would greatly help to establish shared resources such as code repositories, datasets, stories of inadvertent bias (e.g. http://callingbullshit.org/), and lists of best practices. There is a need for leadership in this.

3.2. Evidence, Accountability, and Oversight

While algorithms are being used by a wide variety of people, businesses, and organizations to accomplish different purposes, algorithms should also work for the public good. As an increasingly important and valuable tool, it is important that algorithms are used to improve the public good, or at the least not harm the public good. However, detecting whether this is happening is a difficult problem: *How do we integratively assess the impact of an algorithmic system on the public good?*



Notes from the group discussing "evidence, accountability, and oversight" during the third breakout session.

Impacts may be both positive and negative, and often algorithmic systems will have some of each. It is important to examine different types of impact, and to identify outcomes relevant to the groups that may be impacted by use of each algorithmic system. Impacts can often change as a result of non-algorithms factors such as transparency of processes, government influences, or social norms around use; as such we should strive to evaluate the impacts of algorithms in the context of use.

Impacts are often different for disparate groups of people. Benefits may accrue to some groups and not others; likewise the costs may be higher for some groups than others. It is not enough to evaluate an overall impact without looking at subgroups, and it is not enough to evaluate only specific subgroups; both evaluations are needed to understand impacts.

Impacts cannot be measured by simply looking at the algorithm or its parameters and outputs. Instead, evaluations should be aspirationally causal; we should work to understand how algorithms and decisions based on algorithms cause impacts and then measure the size of those impacts. While causal evaluations (like randomized controlled trials) are not always possible, we should aspire to evaluate the causal impacts rather than simply examining how algorithms work or what parameters are set. We should seek types of evidence that can be used to evaluate causal impacts.

Finally, evaluations of impacts should be related to system characteristics. There is a need for constructive evaluations that don't just shut down algorithms with inappropriate impacts, but instead provide positive contributions that generate solutions.

Why is this difficult? Current attempts to provide oversight and accountability have run into a number of difficult challenges that limit our ability to achieve these goals.

One of the biggest challenges is an incomplete understanding of socio-cultural context by practitioners who use algorithms. Often the impacts that are most relevant to the public good are not direct, intended outcomes of the algorithms, but instead arise out of the context of use. It is unclear what outcomes should be measured in different contexts. Understanding these impacts lies between established disciplines, and requires increased participation across disciplines. This is particularly true uniting technical disciplines with social disciplines.

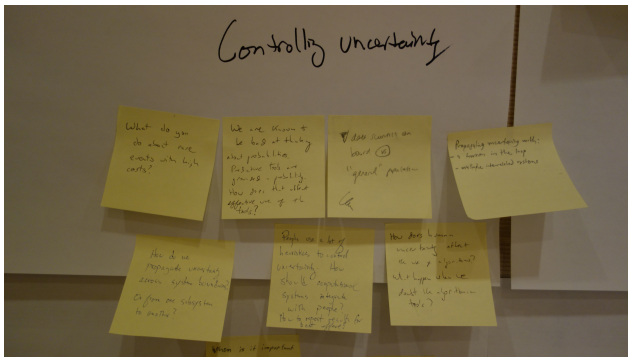
Algorithms and their use in society are creating a lot of societal good. Often the easiest solutions prevent algorithms from being used, or being used effectively; instead we need to focus on evaluating algorithms constructively in ways that preserve the positive characteristics of algorithm use while mitigating the negative impacts. Criticism is not enough, but is an important input.

Recommendations for Progress The growing amount of data and the growing use of algorithms provides many opportunities to better understand algorithm use and to develop methods for evaluating the impacts of algorithms. It would help if algorithms can be designed in ways that produce evidence that can be used to evaluate the impacts of those algorithms. Many current algorithms are black boxes that are difficult to evaluate; producing additional output that can be used as evidence of impacts will greatly assist in evaluating those impacts. For example, algorithms can be designed in ways that support A/B testing and interventions that can be used to evaluate causal impacts.

Researchers across disciplines should look to partner with industry and government organizations that are using algorithms. Increasing public attention to this issue is providing strong incentives for such partnerships, which opens up many opportunities for effective evaluations.

3.3. Handling Uncertainty in Algorithmic Systems

Suppose, as an example, we have a cyber security system that analyzes network traffic for potential threats. It can preprocess the individual packets into feature vectors, try to cluster those vectors into similar groups, and then label those groups with the types of activity that they represent. And analysts then look at these groups and take actions to either allow or prevent the traffic. Each step in this process changes the data in some way, and each step introduces some uncertainty. Preprocessing loses some data and assumes features that may or may not be appropriate; clustering is based on some metric and reduces variation into representative values; and labeling for humans often oversimplifies. Suppose a threat is detected. How certain is the system about the threat? Can we improve the credibility of the detection by incorporating more data?



Notes from the group discussing "uncertainty" after the first breakout session.

There are many sources of uncertainty that arise naturally in the process of using algorithms for decision-making. Uncertainty can arise from data collection methods, from the choice of data to use, from modeling

assumptions, from propagation across subsystems or between analysis steps, from presentation to decision-makers, and in incorporating algorithmic output with other related analyses. *How might we holistically treat and attribute uncertainty throughout data analysis and decision systems?*

Few modern algorithmic systems explicitly acknowledge or formally model the underlying uncertainty in the algorithms' output. One of the biggest challenges in using algorithms for decision-making is attribution: identifying the sources or origination points of uncertainty in algorithmic processes. There are many ways uncertainty can be introduced into an algorithm, and understanding and attributing uncertainty to appropriate sources is essential to understanding algorithm outputs and to making appropriate decisions about their use.

Uncertainty is related to but not the same as errors produced by algorithms. Uncertainty is a lack of knowledge, where errors are mistakes or misclassifications. Uncertainty is answering a question with "I don't know" rather than answering with an incorrect response. Uncertainty and errors are two separate (but related) reasons to trust or not trust an algorithmic system.

Why is this difficult? Uncertainty can be introduced in a large number of places in an algorithmic system. The input data usually has uncertainty (often unacknowledged). The algorithm itself can create and introduce uncertainty. Choices about which data and which algorithm to use can lead to uncertainty. Model choices, such as preprocessing, parameterization, and regularization can all add uncertainty. And the decision context that an algorithm is used in can add substantial uncertainty.

Uncertainty can be propagated across stages of the analysis. Most algorithms assume fixed, accurate inputs, but in reality inputs often have uncertainty that can be compounded across steps in the analysis. Formal modeling of uncertainty is valuable (though rare), and often requires substantially increased computational capabilities to accomplish.

Human beings – both end users of algorithmic outputs and data scientists developing algorithms – are particularly bad at evaluating uncertain information and reasoning about uncertainty. It is difficult for people to incorporate formal uncertainty analyses into thinking and decision-making. People often ignore uncertaint and treat point estimates as facts. These human factors need to be integrated with any formal or informal uncertainty representation to ensure that algorithm outputs are usable and provide value.

One of the biggest challenges that makes this a hard problem is that "the fundamental phenomena do not scale down." That is, uncertainty in large, complex algorithmic systems is not the same as uncertainty in small, simple problems. Uncertainty in something as complex as a self-driving car is difficult to decompose into smaller problems for study, and often must be addressed directly in large, complex systems.

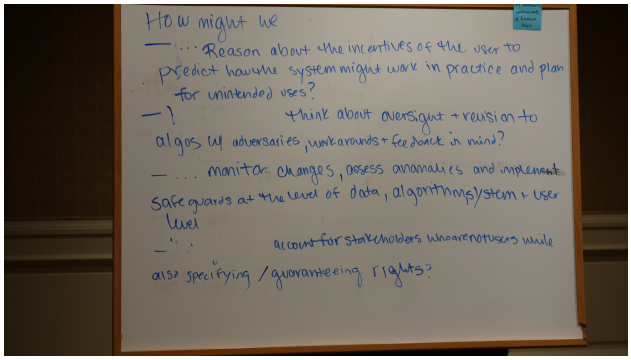
Recommendations for Progress Algorithm designers and data scientists will need increased training in thinking about and working with uncertainty. End users of algorithm outputs will also likely need training, and algorithms will need better ways of representing uncertainty in outputs in ways that humans can understand and reason about. Correct interpretation is not a given, and people may not like uncertain estimates (instead preferring the false certainty of point estimates) even though explicitly uncertain systems may perform better.

We need better ways of representing uncertainty in data, in algorithms, and in the human aspects of algorithmic systems. Right now, uncertainty is represented in very different ways in different domains and for different parts of the system, which makes it very difficult to unify or to accurately represent overall levels of uncertainty in algorithmic outputs. For example, uncertainty can be represented statistically, quantitatively, qualitatively, as a set of discrete hypotheses, or as a guy feeling; and different parts of a system may address it using incompatible representations.

We also need better ways to measure and quantify uncertainty. Current methods are very computationally intense and often require computationally intensive simulation or sampling. These are often impractical on large datasets or for large scale problems where algorithms are being used. Better theory about uncertainty and its relationship with algorithms and decision-making will help. Specifically, we would like to see attempts to unify theories of quantitative uncertainty representations with human theories of decision-making under uncertainty.

3.4. Adversaries, Workarounds, and Feedback Loops

Goodhardt's law states that any useful metric, once it becomes a target, ceases to be useful because people will start to game the metric. The same is true for decisions made based on algorithms; as algorithms are increasingly used to help make important decisions, people will being to game the system. Adversaries – people who are intentionally trying to manipulate the decisions – have many opportunities to influence algorithmic decisions. They can directly alter the choice of algorithm or parameters to the algorithm. They can alter the data used as an input by the algorithm (like Search Engine Optimization seeks to do). They can strategically use or not use a system to affect what data the system has. Or they can influence the interpretation and decision-making based on the algorithm.



Shared notes from the adversaries, workarounds, and feedback loops group in response to the "How might we..." prompt in the third breakout.

Not all manipulation of algorithmic decisions is intentional. Consider, for example, a police system intended to track provenance of evidence that requires all officers to log into a specific computer to log activity. It is easier for one officer to log in and leave the computer logged in. This isn't an intentional subversion of algorithms, but is a workaround that produces distorted outcomes. Another example of an unintentional workaround is judges use of algorithms in sentencing. Evidence has suggested that judges cite algorithm output when it agrees with them, but choose to ignore the output when it disagrees with their preexisting opinion. *How should a trustworthy algorithm account for adversaries, workarounds and feedback loops?*

Why is this difficult? Algorithmic systems are born brittle. Initial versions of algorithms rarely are robust to manipulation. It takes time and use to discover how people will be able to manipulate any algorithmic system. Are there ways we can monitor algorithm outputs and decisions for changes, and identify and assess anomalies as they arise? Currently, we also don't have good methods for implementing safeguards; what would an algorithmic safeguard look like?

Algorithms, and particularly algorithms used to support or make decisions, are one part of a larger system. And it is often the system that is being manipulated. It is important to take a systems perspective that includes not just the algorithm, but the data and the people and the processes and organizations around the algorithm. All parts of the system can be exploited, not just the algorithm.

Part of the systems perspective is including the motivations and incentives of everyone who uses or is affected by an algorithmic system. Understanding what people are trying to achieve and how algorithms alter people's incentives is critical to understanding the human behaviors that may alter or affect the performance of the algorithmic system. It often isn't clear whether a given human is intentionally manipulating the algorithm (an adversary) or unintentionally changing some aspect of the system in response to changing incentives (a workaround), nor whether this distinction is important. Note that this includes both users of the system and non-users, since even people who aren't interacting with a system can still be affected by it and influence it.

Recommendations for Progress When algorithms are in use in practical settings (not in the lab), it is important to continually monitor the algorithms and inspect them for evidence of manipulation. However, most algorithms right now are difficult to inspect or understand. We need to develop better methods of inspecting and monitoring the performance of algorithms, and methods for identifying anomalous outputs or biased patterns of output. To do this, we need to be better at identifying clear goals for algorithm output and metrics and measures for those goals.

As algorithms are incorporated into a wider variety of systems and processes, we need to remember that it is the system as a whole that is often the target of attack. Metrics for the performance and bias of algorithmic systems should include whole system metrics, including the expected properties of input data

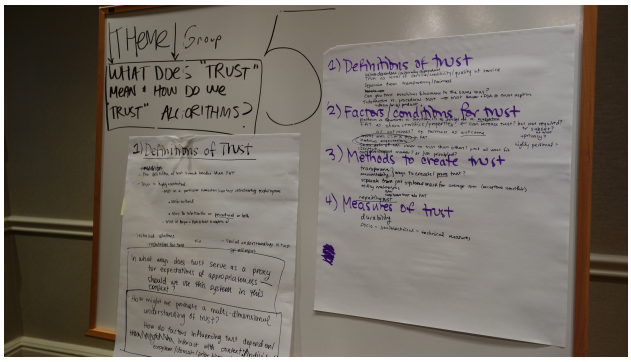
and the performance of humans involved with the system. The human factors of algorithmic use are often an important aspect of vulnerability to adversaries and workarounds.

It would really help to clarify possible adversary threat models – including workarounds as potential threats. We recommend developing a taxonomy of adversarial behaviors in practice around algorithmic systems, to help algorithm designers better understand them.

Finally, we recommend increased effort to design algorithmic systems robust to different types of manipulation. Most advanced algorithms are initially brittle and vulnerable to many forms of manipulation; but with some effort we believe algorithms can be designed that perform well and are either robust to manipulation or that make attempted manipulation evident.

3.5. How Do We Trust Algorithms?

Algorithms are not used in isolation, but instead are usually components of some larger system (like a self-driving car). Algorithms are chosen, calibrated and developed as part of that system, and there are a large number of processes around the use of the algorithm. Trust in algorithms often arises not from the technical properties of the algorithm, but from the people and processes that make decisions about its use. *What are the processes through which different stakeholders come to trust an algorithm?*



Shared notes about trusting algorithms in the third breakout.

Processes can produce trust in algorithms and in systems that use algorithms in multiple ways. Substantive trust arises directly from the algorithm, and includes things like using appropriate weights, satisfying fairness properties, and being transparent. Procedural trust is also needed, and come from the processes and people around the use of the algorithm such as who created the algorithm, who chose which algorithm to use, and how those choices are made. Trust can be socially derived when trusted peers vouch for a system or experientially derives as an individual uses or interacts with an algorithmic system. Research is needed to improve all of these forms of trust.

To improve the trustworthiness of algorithmic systems, it is important to acknowledge and study the range of stakeholders involved in using these algorithms. The creators of algorithms have important roles in identifying properties of algorithms, creating algorithms with valuable and trustworthy properties, and communicating clearly with other stakeholders. Data providers play an important role in the trustworthiness of algorithmic systems, and it is important to understand how data providers can help or harm trust. Algorithms used for decision-making have a wide variety of users and people who are affected by the decisions. There are also often regulators, community members, and other participants in markets that are affected by how algorithms are used.

As an interesting case study, a group from the New York City Association for Child Services with present at the workshop. They have developed an algorithmic decision aid for use in their organization, and described how they built trust with the wide range of stakeholders including children, families, the NYC city council, agency staff, and the city as whole. Rather than buying a commercial algorithmic system, they built their own. "We didnt buy our algorithm from a black-box vendor. We know exactly what went into the soup of our predictive analytics models." They also created an advisory group with representatives from the different groups of stakeholders that assisted with decision-making, trained the entire data team in ethical use of algorithms, and ensured everyone involved had appropriate credentials and credibility.

Why is this important? Algorithms are increasingly being used to make decisions that affect people. Real people can be negatively impacted by decisions made by algorithms or with the assistance of algorithms. Increasingly, algorithms are being used in more consequential situations. Algorithms have the potential to improve these decisions, but only if they are trusted. A lack of trust in the algorithms or in the algorithmic systems prevents the system from realizing its potential, limits its usefulness, and can cause backlash against related systems.

Why this is difficult? Trust is often irreducible. Trust is needed in the overall system, not just in individual components. Measuring trust in individual components is not enough. Furthermore, there are a large number of individual components and processes that go into algorithms and algorithmic systems that create separate pathways to trust. Trustworthiness does not necessarily transfer from subcomponents; just because part of an algorithmic system is trusted does not mean the whole system will be.

Trust is not a static property of an algorithmic system. Rather, trust builds over time due to processes and experiences. Trust is not durable; it can changed, it can be broken, and it can (sometimes) be repaired. However, the timescales for building trust in systems are often outside of normal laboratory experiments.

Recommendations for Progress Trust in an algorithm or a system that used algorithms cannot be reduced to simple algorithm metrics like fairness or bias. It would help to better understand how properties of algorithms are interpreted and used to build trust in the larger systems that use those algorithms. There are many socio-technical factors involved in trusting an algorithmic system that need to be studied in addition to technical properties of algorithms.

Trust is built over time, and more research needs to be conducted about how trust is built or lost in algorithmic systems rather than simply assuming that trust is a static property of the system or algorithm. Is it possible to build algorithms that can rebuild trust after a breach in trust?

Understanding the trustworthiness of algorithms requires access to internal processes and deliberations of stakeholders. However, there are incentives to keep these processes secret to avoid opening up the technology. It would help if there are positive incentives for sharing tools, techniques, and processes that enable trust to be built in algorithms. Can we create such incentives? Can we support the sharing of best practices?

4. Additional Insights and Recommendations

In the group discussions about these challenges, some additional cross-cutting themes and issues emerged that the group thought were important.

Algorithm use, trust in algorithms, and trust in decisions made using algorithms is very contextual. It is important to study these algorithms in real contexts of use to understand the wide range of influences on trust and effects that trust has in algorithms. Evaluation in context is also important for understanding how real adversaries can attack the system, how real users work around the algorithms, how people deal with the uncertainty in the algorithm, and who makes decisions about algorithm and how their training and abilities influence algorithm use and trustworthiness.

Systems that use algorithms for decision-making are inherently socio-technical. They include important technical features and capabilities; they include people who are using algorithms, making decisions about their use, and who are affected by the algorithms; and they include complex interactions between the technology and people that produce the outcomes. We need more research into how the technical and human parts of algorithm use combine to produce either trustworthy or untrustworthy systems.

It is difficult to determine appropriate comparison cases when evaluating algorithmic systems, and as a result algorithms are often being held to unrealistic standards. We need to develop better methods for identifying and measuring the counter-factual best alternative to an algorithm to determine more accurately the effects of algorithm use and the benefits, downsides, and biases of algorithms. We also need better metrics for quantifying risk in algorithmic systems and bias in algorithmic systems.

Discussions of trustworthiness always lead to discussions of governance mechanisms that lead to trust. Governance is complicated, and too many proposed governance solutions are impractical oversimplifications. More research is needed on how algorithmic systems can be governed and what is needed in terms of people processes and properties of algorithms for this governance to be successful.

There are important resource limitations that limit our ability to study many issues around algorithmic decisions. Many algorithms simply don't work on small scales that individual researchers have access to. It is important to create large teams with strong ideas, and to engage in "collective bargaining" on behalf of the research community to gain access to datasets and in-use algorithms for study.

Finally, the issues brought up in this workshop and report are frequently international in scope, and there is broad interest across a large number of nations in these issues. We need to find better ways to collaborate with colleagues in all nations to work on and address these issues.

References

[1] Mike Ananny and Kate Crawford. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 33(4):1–17, 2017.

[2] J.D. Biersdorfer. How waze tries to keep its crowd honest, 2016. URL https://www.nytimes.com/2016/09/17/technology/personaltech/how-waze-tries-to-keep-its-crowd-honest.html.

[3] Kelsey Gee. In unilever's radical hiring experiment, resumes are out, algorithms are in, 2017. URL https://www.wsj.com/articles/in-unilevers-radical-hiring-experiment-resumes-are-out-algorithms-are-in-1498478400.

[4] Andrew J. Hawkins. Delta air lines plans to use facial recognition to speed up bag drops, 2017. URL https://www.theverge.com/2017/5/15/15640568/delta-facial-recognition-self-service-bag-drop-minneapolis.

[5] K A Hoff and M Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 57(3):407–434, 2015.

[6] Bernie Hogan. From invisible algorithms to interactive affordances: Data after the ideology of machine learning. In *Roles, Trust, and Reputation in Social Media Knowledge Markets*, pages 103–117. Springer International Publishing, 2015.

[7] John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004.

[8] Joseph CR Licklider. Man-computer symbiosis. *IRE transactions on human factors in electronics*, 1: 4–11, 1960.

[9] Farhad Manjoo. Can facebook fix its own worst bug?, 2017. URL https://www.nytimes.com/2017/04/25/magazine/can-facebook-fix-its-own-worst-bug.html.

[10] Frank Pasquale. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press, 2015.

[11] Samantha Schmidt. A massive earthquake was reported in california wednesday — by mistake, 2017. URL https://www.washingtonpost.com/news/morning-mix/wp/2017/06/22/a-massive-earthquake-was-just-reported-in-california-turns-out-it-happened-in-1925/?utm_term=.d84a1da8f0a0.

[12] Jack Stilgoe. What will happen when a self-driving car kills a bystander?, 2017. URL https://www.theguardian.com/science/political-science/2017/jun/24/what-will-happen-when-a-self-driving-car-kills-a-bystander.

[13] Jason Tashea. Courts are using ai to sentence criminals. that must stop now., 2017. URL https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stop-now/.

[14] Jerry Useem. How online shopping makes suckers of us all, 2017. URL https://www.theatlantic.com/magazine/archive/2017/05/how-online-shopping-makes-suckers-of-us-all/521448/.

Workshop on Trustworthy Algorithmic Decision-Making

trustworthy-algorithms.org

Day 1: Monday, December 4	
7:30 AM — 8:30 AM	BREAKFAST (buffet)
8:30 AM — 8:45 AM	Opening Remarks
8:45 AM — 9:00 AM	Lightning Introductions *30 seconds each: tell us your name, affiliation, and one sentence summarizing your interests related to the workshop.*
9:00 AM — 10:30 AM	Panel Session + Discussion 1 • David Weinberger, Harvard University • Rishi Ahuja, Consumer Financial Protection Bureau • Stacy Wood, University of Pittsburgh • Bo Cowgill, Columbia University • Vincent Conitzer, Duke University • Min Kyung Lee, Carnegie Mellon University
10:30 AM — 10:45 AM	Coffee and Snack Break
10:45 AM — 12:15 PM	Panel Session + Discussion 2 • Joshua Kroll, University of California, Berkeley • Maria Rodriguez, Hunter College, City University of New York • Ling Liu, Georgia Institute of Technology • Chuck Howell, MITRE Corporation • Inbal Talgam-Cohen, Hebrew University of Jerusalem • Shion Guha, Marquette University
12:15 PM — 1:45 PM	LUNCH (buffet) and Affinity Diagramming
1:45 PM — 2:00 PM	Breakout Instructions and Questions
2:00 PM — 3:15 PM	Parallel Breakouts | Stage 1: Brainstorm *Look on the back of your badge for your assigned breakout room.*
3:15 PM — 3:30 PM	Coffee and Snack Break
3:30 PM — 4:45 PM	Parallel Breakouts | Stage 2: Synthesize *Look on the back of your badge for your assigned breakout room.*
4:45 PM — 5:00 PM	Reflections on Day 1 *What's one thing you thought about today you haven't thought about before related to trustworthy algorithmic decision-making?*
6:30 PM — ???	Workshop Dinner @ Fyve

Workshop on Trustworthy Algorithmic Decision-Making

trustworthy-algorithms.org

Day 2: Tuesday, December 5	
7:30 AM — 8:30 AM	BREAKFAST (buffet)
8:30 AM — 8:45 AM	Opening Remarks
8:45 AM — 10:00 AM	Parallel Breakouts | Stage 3: “How Might We…” *Look on the back of your badge for your assigned breakout room.*
10:00 AM — 10:15 AM	Coffee and Snack Break
10:15 AM — 12:00 PM	Parallel Breakouts | Stage 4: Problem Statement *Look on the back of your badge for your assigned breakout room.*
12:00 PM — 1:00 PM	LUNCH (buffet)
1:00 PM — 2:15 PM	Problem Statement Presentations, 15-20 minutes each
2:15 PM — 3:15 PM	Plenary Discussion: Structuring the Problem Space
3:15 PM — 3:30 PM	Closing Remarks
3:30 PM — 3:45 PM	Coffee and Snack Break
3:45 PM — ???	Integration, Synthesis, and Initial Drafting

Overview of Breakout Sessions

The purpose of this workshop is to develop ideas that will further define the problem space, the key problems and the critical questions that need to be answered to make progress toward understanding, developing, and evaluating of Trustworthy Algorithmic Decision-Making.

There will be 4-5 parallel sessions during each time set aside for breakouts on the agenda. Each parallel session will focus on one of the high-level themes that emerge from the note-taking and affinity diagramming during the first half of Day 1. Workshop participants will rotate through the different themes, working on a different theme during each scheduled breakout session. Each breakout session has a specific focus: Brainstorm, Synthesize, "How might we…", and Problem Statement, such that a small group will be working on each theme during each stage, and then hand off their work to the group working on that theme during the next stage. At each stage, the groups will be randomized so that everyone gets to meet, work with and bounce ideas off of new people.

Each parallel breakout session is 1 hour and 15 minutes long; the last 15 minutes should be spent capturing and documenting for the next group. Don't forget to do quick introductions first thing during each breakout session!

Roles and Responsibilities

- Theme Champion: one volunteer who stays with a theme through all four stages. Provides continuity by answering questions about earlier conversations. Responsible for keeping the discussions focused and on track, overseeing documentation of the work during each breakout session so the next group can build on what the previous group did, and collecting any files or photos that were taken and storing them in the location provided for each theme. (Ignore the room assignments on the back of your badge if you are a Theme Champion!)
- Note-Taker(s): PhD student participant who stays with a theme through all four stages of the parallel breakouts. Any additional participants who want to help with note-taking can also do so. Responsible for documenting the work during each breakout session, including notes and photos of post-its and anything that is written up on the whiteboard; this will all be invaluable for writing the report.
- Time-Keeper: Chosen at the beginning of each breakout session. Responsible for making sure the group stays on task, and stops ~15 minutes before the end of the breakout session to wrap up and document the work.
- Participants: Responsibilities include… participating! Keep an open mind, be inclusive, remember that the group is diverse and ask for questions and clarifications when necessary. If you have a question someone else probably does too! Be creative and patient, and have fun!

Parallel Breakouts | Stage 1: Brainstorm

The **goal** of this activity is to creatively generate ideas and background information to add content and context and further develop the theme. This is an expansion phase, not a reduction phase. The **main output** of this phase is the documented ideas that the group generates.

- Start by developing a question or prompt for the brainstorming that characterizes the theme, based on the group affinity diagram.
- Then do three short rounds of brainstorming, 10-15 minutes each, in response to the question or prompt, writing each idea on a post-it. Write first, then share later.
- After each round of brainstorming, each person sticks their post-its on the wall/whiteboard and reads/ describes it. Listen to each other, and in the next round build on each others' ideas!
- Aim for quantity! Come up with as many ideas as possible. Encourage weird and wacky ideas.
- Stay in a generative mindset, not a critical one. Keep an open mind, and be positive. One way to do this is to encourage "and" statements, not "but" statements.

Parallel Breakouts | Stage 2: Synthesize

The **goal** of this activity is to build on the idea generation in the previous phase, and identify the big ideas and key concepts related to the overarching theme. The **main output** of this phase is at least 3-5 "insight statements" about problems that need to be understood better and/or solved, along with text to describe each insight.

- Start by walking the wall for 10-15 minutes, and reading the ideas that the previous group generated. Add post-its if you have new ideas, observations, reactions, etc.
- Then discuss with the group what's on the wall, and identify at least 3-5 "big ideas".
- Write an "insight statement" for each big idea. To do this, discuss each idea, and rephrase it as a short statement that captures an understanding that sheds light on some important aspect of the theme. This doesn't need to be perfect; it is just a building block for the next stage.
- Write some text to describe each insight, and refine. Borrow heavily on the output of the brainstorming and ideas from the group affinity diagram.

Parallel Breakouts | Stage 3: "How Might We…"

The **goal** of this activity is to expand on the insight statements, and rephrase them as questions that need to be answered. This transforms the thinking about the insights into opportunities for future research and design activities. The **main output** of this phase is one question per insight statement, along with notes captured from the discussion.

- Start by reading the insight statements and supporting text generated by the previous group.
- Rephrase the insight statements as questions that need to be answered, starting with "How might we…". The questions should be broad enough to allow for a variety of possible approaches and answered, but narrow enough that they are not overly restrictive
- For each "How might we…" question, discuss and take notes on: examples in the world (domains?), approaches/methodologies, constraints, stakeholders, leverage points, etc.
- Write each idea up on the whiteboard or a giant post-it to hand off to the next stage.

Parallel Breakouts | Stage 4: Problem Statement

The **goal** of this activity is to select one problem statement from the candidates produced in the previous session, and further describe it. The **main output** is a presentation about it that you will deliver to all of the workshop participants after lunch on Day 2.

- Start by reading and discussing the output of the previous stage, which should consist of ideas written on the whiteboard or giant post-its, capturing new ideas as they come up.
- Choose one that you will focus your presentation on. We recommend using dot-voting (aka sticker voting, multi-voting, etc.) so that the choice is not dominated by individual voices in the room. 3-4 votes each should be enough although you may need more if there are a large number of candidates. Vote by putting one or more dots next to your favorite idea.
- Once you have selected an idea, write a problem statement that is ambitious, but still actionable, and in line with the goals of the workshop.
- Then work on your presentation. Remember to avoid jargon, and define discipline-specific terms.

The presentation should cover the following:

1. What is the problem and why is it important?
   - define key terms and identify stakeholders
   - provide a scenario/example that illustrates the problem
   - what are the best sources of information about the problem?
2. Why is this a difficult problem?
   - describe the scope of the problem
   - what are the important unsolved/poorly specified aspects?
3. Why is progress possible?
   - describe what progress would look like; how would we recognize it?
   - approaches likely to make progress
4. What are the barriers for success, and how might we mitigate them?
   - ideas, training, incentives, resources (time, funding, data, etc.)…