People + AI Research 2019 Symposium Reader

PAIR

People + AI
Research

People + AI Research 2019 Symposium Reader

Introduction

Selections from the People + AI Guidebook

A toolkit of methods and best practices for designing human-centered AI experiences

Responsible by Design

Introduction

Welcome to the 2019 PAIR Symposium.

Inspired by the participatory design movement, this year's symposium theme Participatory Machine Learning is an approach to building machine learning systems that actively involve a diversity of stakeholders – technologists, UXers, policymakers, end users, citizens.

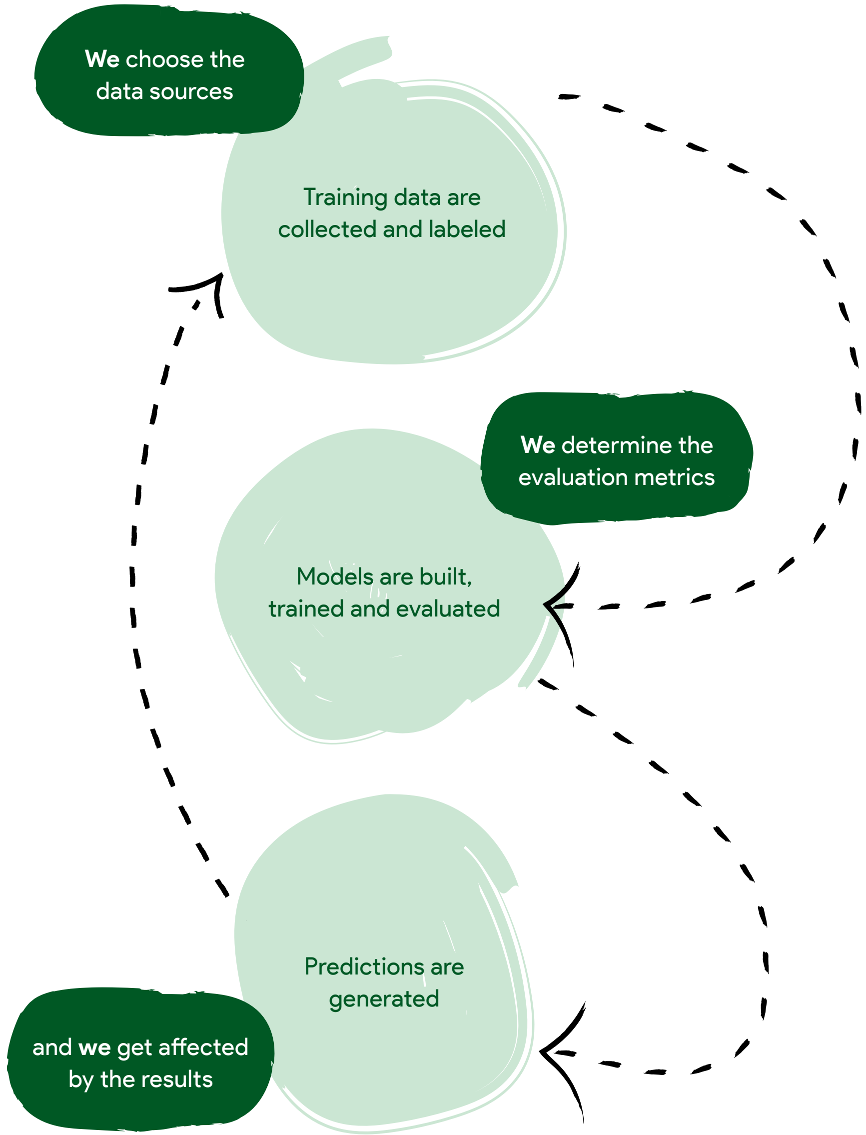
This isn't a new story: I personally saw the value of making technology more participatory as I worked on data visualization throughout my career. When I first entered the field, visualization was a tool largely for the elite in academia and business. But today sophisticated visualizations tell stories ranging from investigative journalism to artistic portraiture. The key to that transformation was creating tools and systems that made visualization accessible to a much broader community – and, most important, finding ways to let that community shape the development of new technology. I believe the time has come for a similar transformation in how to build and deploy AI.

In this symposium guide you will find excerpts from the recently released People + AI Guidebook as well as reflections on the intersection of design and policy from our writer in residence David Weinberger. Indeed, in some sense design may be viewed as *applied policy*, since every design decision – consciously or not – ultimately carries policy implications.

We are excited to spend the day discussing how we all can play a role in improving the design, use and evaluation of AI systems. In the spirit of participation, I look forward to seeing what new collaboration opportunities emerge from this gathering.

Fernanda Viégas and the PAIR Team

Human perception drives virtually every facet of machine learning



PAIR is devoted to advancing the research and design of user-centric AI systems. We're interested in the full spectrum of human interaction with machine intelligence, from supporting engineers to understanding everyday experiences with AI.

Our goal is to do fundamental research, invent new technology, and create frameworks for design in order to drive a human-centered approach to artificial intelligence. And we want to be as open as possible: we're building open source tools that everyone can use, hosting public events, and supporting academics in advancing the state of the art.

To learn more about PAIR's work, you can find us at:
PAIR research & tools: **ai.google/pair**
People + AI Guidebook: **pair.withgoogle.com**

Selections from the People + AI Guidebook:

A toolkit of methods and best practices for designing human-centered AI experiences.

User Needs + Defining Success

Identify user needs, find AI opportunities, and design your reward function.

Aligning your product with user needs is step one in any successful AI product. Talking to users, looking through data, and observing behaviors can shift your thinking from technology-first to users-first.

Which user problems is AI uniquely positioned to solve?

Find the intersection of user needs & AI strengths. Make sure you're solving a real problem in a way where AI is adding unique value. When deciding on which problem to solve, you should always build and use AI in responsible ways. Take a look at the Google AI Principles[1] and Responsible AI Practices[2] for practical steps to ensure you are building with the greater good in mind.

How can we augment human capabilities in addition to automating tasks?

Automate tasks that are difficult or unpleasant, and ideally ones where users who do it currently can agree on the correct way to do it. Augment bigger processes that people enjoy doing or that carry social value.

How can we ensure our reward function optimizes AI for the right thing?

The *reward function* is how an AI system defines successes and failures. You'll want to deliberately design this function including optimizing for long-term user benefits by imagining the downstream effects of your product and limiting their potentially negative outcomes.

[1] Google AI Principles: ai.google / principles

[2] Responsible AI Practices: ai.google / responsibilities / responsible-ai-practices

Data Collection + Evaluation

Decide what data are required to meet your user needs, source data, and tune your AI.

Data is the bedrock of any ML system. Having responsibly sourced data, from a relevant context, checked for problematic bias will help you build better systems and therefore more effectively address user needs.

Does the training dataset have the features and breadth to ensure our AI meets our users' needs?

Think carefully about what features, labels, and examples you will need to train an effective AI model. Work systematically to break down user needs, user actions, and ML predictions into the necessary datasets. As you identify potential datasets, or formulate a plan to collect them, you'll need to be diligent about inspecting the data, identifying potential bias sources, and designing the data collection methods.

Once you have a model, you will need to test and tune it rigorously. The tuning phase involves not only adjusting the parameters of your model, but also inspecting your data – in many cases, output errors can be traced to problems in your data.

Should we use an existing training dataset or develop our own?

As part of sourcing data, you'll need to consider relevance, fairness, privacy and security. You can find more information in Google's AI Principles and Responsible AI Practices. These apply whether you are using an existing dataset or building a new training dataset.

How can we ensure that raters aren't injecting error or bias into datasets when generating labels?

Correctly labeled data is a crucial ingredient to an effective supervised ML system. Thoughtful consideration of your raters and the tools they'll be using will help ensure your labels are accurate.

Mental Models

Introduce users to the AI system and set expectations for system-change over time.

Because AI-powered products can adapt and get better over time, the user experience can change. Users need to be prepared for that, and adjust their mental models as necessary.

Which aspects of the AI system should we explain to our users?

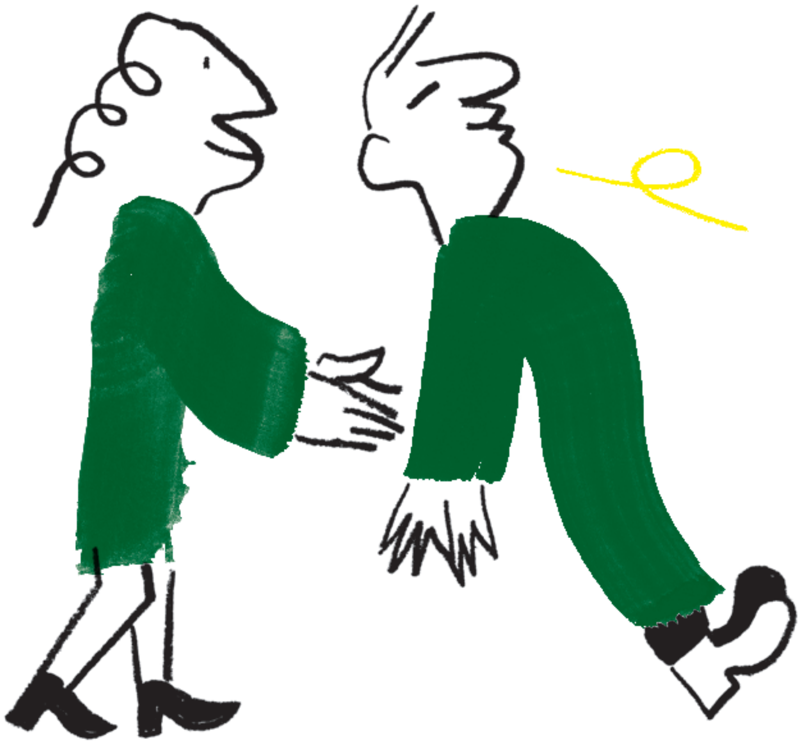
Set expectations for adaptation. Help people get the most out of new AI uses by identifying and building on existing mental models. Ask yourself questions like *What is the user trying to do?*, *What mental models might already be in place?*, and *Does this product break any intuitive patterns of cause and effect?*

How should we introduce AI to the user initially – and thereafter?

Set realistic expectations early. Describe user benefits, not technology. Describe the core value initially, but introduce new features as they are used. Make it easy for users to experiment with the AI in your product.

Plan for co-learning. Connect feedback to personalization and adaptation to establish the relationship between user actions and the AI output. Fail gracefully to non-AI options when needed.

Explainability + Trust



Explain the AI system and determine if, when, and how to show model confidence.

Explaining an AI system's output can be critical for building trust. The choice of if and how to offer an explanation needs to be made with the user in mind to ensure that it is intelligible and suitably framed.

How much should the user trust the AI system?

Help users calibrate their trust. The goal of the system should be for the user to trust it in some situations, but to double-check it when needed. Factors influencing calibrated trust are:

- **Articulate data sources** Telling the user what data are being used in the AI's prediction can help your product avoid contextual surprises and privacy suspicion and help the user know when to apply their own judgment.
- **Tie explanations to user actions** Showing clear cause-effect relationships between user actions and system outputs with explanations can help users develop the right level of trust over time.
- **Account for situational stakes** Providing detailed explanations, prompting the user to check the output in low-confidence / high-stakes situations, and revealing the rationale behind high-confidence predictions can bolster user trust.

How should we show users the confidence associated with an AI prediction?

When a user needs to make a decision based on model output, when and how you display model confidence can play a role in what action they take. There are multiple ways to communicate model confidence, each with its own tradeoffs and considerations.

Feedback + Control

Design feedback and control mechanisms to improve your AI and the user experience.

Feedback and user control are critical to developing communication and trust between your user and the system, and for developing a product that fulfills your users' needs consistently over time.

How should the AI system request and respond to user feedback?

Communicate value & time to impact. Understanding why people give feedback, and building on existing mental models to explain benefits and communicate how user feedback will change their experience, and when.

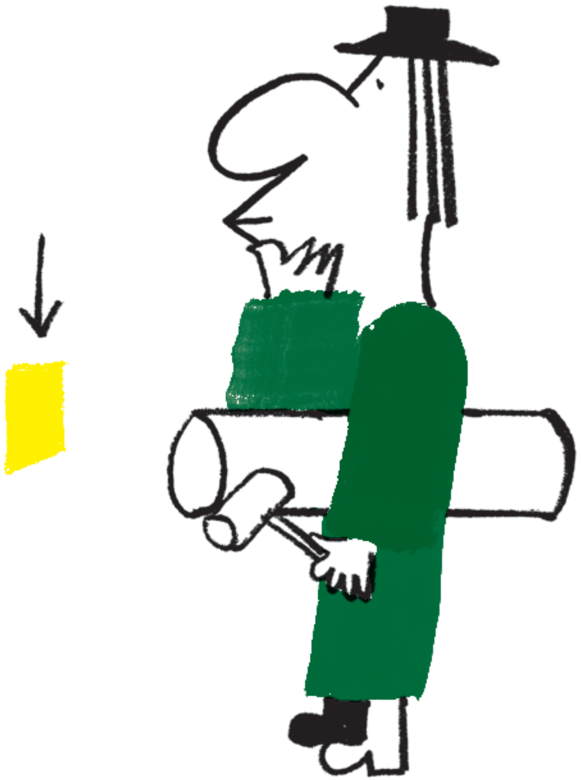
How can we ensure our AI system can interpret and use both implicit and explicit user feedback?

Align feedback with model improvement. Clarifying the differences between implicit and explicit feedback, and asking the right questions at the right level of detail.

What's the right level of control and customization to give our users?

Balance control & automation. Helping users control the aspects of the experience they want to, as well as easily opting out of giving feedback.

Errors + Graceful Failure



Identify and diagnose AI and context errors and communicate the way forward.

When designing your error experience, be human, not machine. Address mistakes with humanity and humility, and explain the system's limits while inviting users to continue forward.

When do users consider low-confidence predictions to be an *error*?

When dealing with a probabilistic, dynamic system, a user could perceive a failure in situations where the system is working as intended. Acknowledging that a product is a work-in-progress can help encourage the adoption and feedback that designers and engineers need to continue improving the AI system.

How will we reliably identify sources of error in a complex AI system?

The inherent complexity of AI-powered systems can make identifying the source of an error challenging. It's important to discuss as a team how you'll discover errors and discern their sources.

Does our AI system allow users to move forward after an AI failure?

No matter how hard you work to ensure a well-functioning system, AI systems are probabilistic by nature, and like all systems, will fail at some point. When this happens, the product needs to provide ways for the user to continue their task and to help the AI system improve.

Guidebook recommendations are based on data and insights from Google product teams and academic research. As our research continues to evolve, so will the way we approach AI product design. Expect updates to the Guidebook as the field of AI moves forward – think of it as a living document.

Visit **pair.withgoogle.com** for the entire People + AI Guidebook

Responsible by Design – AI design's policy dimension

By David Weinberger, Ph.D.

Design always expresses, implements, and reinforces decisions about values. Design thus always intersects policy considerations.

This is especially the case for machine learning (ML) given its newness, novelty, and importance. Indeed, ML systems should be Responsible by Design.

All human products have policy implications, but this is especially important when it comes to the sort of AI known as machine learning because of what makes ML distinctive.

Here are seven foundational aspects of machine learning and how they intersect policy concerns.

Participatory Responsible Design

A Responsible by Design process assumes Participatory Design as an important element, for responsible behavior means acting with consideration of the interests of everyone affected.

Assuming those interests without actually involving those whose interests are at stake is not only disempowering, it is very likely to miss interests, misunderstand them, or misjudge their importance.

Participatory Design is important for the development of systems that are responsible by being responsive to the genuine needs of all stakeholders.

ML Learns from Data

The defining characteristic of machine learning is that it develops its own models – a representation of the elements and dynamics of a domain – by statistically analyzing data.

Traditional computing, on the other hand, starts with a model developed by a programmer who has identified the relevant factors and the relationships among them. For example, a traditional weather prediction program begins with a developer working with meteorologists to identify the factors that affect the weather – air temperature, moisture, wind speeds, etc. – and how those factors interact (e.g., when moist air hits cold air, expect precipitation).

Machine learning, on the other hand, starts with data and virtually no prior model.[1] By iterating on the data – numbers shorn of the meaning they have for humans – the system finds correlations of differing strengths among what may be millions of data points. In the sort of machine learning known as deep learning, those correlations may be instantiated as an artificial neural network in which massive numbers of points are related to massive numbers of other points, computed into multiple *layers* that ultimately result in the system classifying the data sufficiently accurately. (What counts as sufficient is decided by the people for whom the system is being built.)

The fact that ML models are constructed directly from data and not from a set of principles or generalizations (e.g. *cold air can cause precipitation when it encounters warm, moist air, cats have pointy ears, the number eight consists of two circles stacked vertically*) inverts how we generally think about policy which by its nature provides general rules or principles that are then applied to particular cases. These two approaches to the same set of concerns can make the conversations among designers and policy-makers especially rich.

ML Can Learn The Biases Implicit in Data

Learning from data may sound like an evidence-based way to avoid human biases, but as has become well known, it can all too easily be an avenue for the injection of human bias. This is for two reasons:

First, if data is about some aspect of human society that includes biases, then – unless care is taken – it may well reflect those biases. For example, an ML system might *learn* that there is a poor correlation between being a woman and being a senior manager because the training data reflects a societal bias against putting women into senior management positions.[2] Or the ML system might *learn* that having a certain skin color makes one a poor risk for a loan, reflecting another societal bias, and so on.

Second, because biases reach deep into culture, they can correlate with unexpected factors. For example, even if a data set contains no information about income levels, factors such as geographic location, health records, grocery purchases, and taste in music might conceivably

① *Virtually* because the choice of the data from which the system will learn usually has an implied model that determines which factors are considered relevant. There are also strong arguments for adding causal models to at least some machine learning models.

② For an interesting discussion of the risks of bias in job application systems, see Ifeoma Ajunwa, 'Beware of Automated Hiring,' New York Times, Oct. 8, 2019 bit.ly/ajunwaHiring serve as proxies for income. It can be difficult or impossible to anticipate all such proxies.

The growing awareness of machine learning's vulnerability to biases has led ML design into areas traditionally the realm of policy makers: What are the biases implicit in our culture? How do these biases manifest themselves? How can we limit their effect? If this inspires bringing in people from other fields that also deal with these issues – historians, philosophers, social scientists, community activists, poets – so much the better. AI designers can especially learn from policy makers' thoughtful and focused ways of thinking about such questions.

AI can also have the salutary effect of discovering unexpected correlations and ways in which pernicious biases are embedded in the culture and society …a first step in rooting them out.

Scanned digit

Digit	0	1	2	3	4	5	6	7	8	9
Probability	1	1	6	16	3	60	3	5	4	1

ML is Probalistic

When you input a scan of a handwritten digit into a machine learning system that has been trained on thousands of scanned digits, it will come up with a probability for each of the ten options:

Even if the system is designed to give the end user a single answer – *It's a 5* – internally its confidence is always probabilistic.

Because ML's statistical analyses result in probabilistic conclusions, ML systems can always be wrong. As the *People + AI Guidebook* makes clear, this has implications for how the output of ML systems should be presented to end users. For example, the system's exact level of confidence in its response probably doesn't need to be made explicit if it's playing a game of chess against you, but might be if it's presenting a medical diagnosis.

But the *Guidebook* also makes clear that the implications go far beyond that.

For one thing, ML's probabilistic nature means that errors are an expected part of even the most successful ML systems. For example, if an ML system can sort chicken eggs by sex with 98% confidence, then if it's wrong 2% of the time, it is a successful system.

In other instances, however, it may be necessary to provide processes for correcting errors and redressing those who suffer from a system's errors, even if those errors are within the predicted range. For example, in the United States, if a switch to autonomous vehicles (AV) were to bring a 90% reduction in traffic fatalities, that would still leave 3,600 people killed by AVs every year.[9] The inevitability of error means that such a system should consider providing ways for people to correct or report errors, continuously monitor error rates, and have mechanisms in place to address those for whom the system fails.

It also means that a Responsible by Design system should be able to distinguish between the inevitable errors a probabilistic system will make, and errors caused by the sort of mistakes and oversights that all human-produced systems are heir to.

ML Requires Exquisite Explicitness

The way ML works requires designers to confront questions with an explicitness that traditional computing can sometimes escape. Often this requires thinking about what otherwise would be identified as questions of policy.

[9] This is the figure often used. E.g., Adrienne LaFrance, 'Self-Driving Cars Could Save 300,000 Lives Per Decade in America', The Atlantic, Sept. 29, 2015 bit.ly/lafranceStat

There is a small irony here, for one of the great strengths of ML is that the developers do not have to explicitly specify all the elements and rules of the domain. Instead, they feed in data, not the logic of how we think those data go together.

But the developers do have to know precisely what will count as a successfully trained system – the reward, loss, or optimization functions. Machine learning systems in fact give developers very fine control over exactly what is wanted from the system.

For example, if Acme Mortgage Loans has your team designing an ML system that will process loan applications to find the one hundred best applicants – regulations permitting – Acme is going to have to tell you exactly what it counts as *best*. The set of tools that machine learning puts in the designer's hands enables the design team to play what-if with complex models the way spreadsheets enable people to try out alternatives, although the logic of spreadsheets tends to be far, far less complex than the logic expressed by neural networks.

So, what will Acme decide to count as a successful allocation of loan grants? Should they simply go with the hundred people most likely to pay back their mortgage loans? Is success the mix of risk and loan size that's likely to make the most money for Acme? Might it be a mix that will lower Acme's revenue projections but also lower the chance of a catastrophic series of defaults? Or perhaps Acme is willing to tolerate higher risk for relatively small loan amounts because the company has a principled commitment to enabling lower income families to buy their first homes.

The discussion of what constitutes success goes yet deeper than that, for Acme and the society in which Acme operates may have a profound interest in making sure such decisions – made by people alone or with the assistance of AI – are fair. And Acme may discover that some of the most desirable business options result in what at least seem like unfair results overall. For example, if thirty percent of the applicants are women, Acme might find that under some definitions of success, only ten percent of the loans recommended by the ML go to women applicants. This should occasion deliberations about what exactly counts as fair.

For example, some at Acme might argue that so long as gender and its proxies are not affecting the outcomes, the decisions were *gender blind* and therefore fair, no matter what percentage of the loans go to women.

(This might well occasion an investigation of the data and the model to look for hidden biases.)

Others might insist that it's only fair if about thirty percent of loans go to women, roughly matching the percentage of female applicants, even if that requires lowering the threshold of the risk of default for women.

Others might argue that the percentage of women recipients should match the national demographic of 51%.

Others might say that it's fair if the percentage of false positives (people granted loans who turn out not to pay them back) is the same for men and women. Likewise of the percentage of false negatives (people denied loans who should have been approved) is the same for both genders.

And others might question the assumption, quite possibly baked into the data set, that gender is binary.

These are decisions that require many areas of expertise, many skills, and, most important, many points of view. The Responsible by Design approach entails having a wide, diverse team that includes not only the organization sponsoring the creation of the ML system and the team of engineers, UX designers, and others who are developing it, but also those from the communities of people who will be affected directly or indirectly by its deployment.

The technical design team of course brings its own talents to the discussion. Machine learning's picky literalism has required computer and social scientists to take what seemed like a simple concept – fairness – and to think clearly about its many varieties, some of which are easier to understand if one has an understanding of how machine learning builds models. The technical designers can help the rest of the team understand the varieties of fairness available, and enable Acme to to make an informed choice, perhaps by having *what-if* tools built into the system.

The specificity required by ML systems may also be useful to policy makers. Dialogue between AI designers and policy makers can help inform both sides, and advance the development of options and ideas.

ML Requires Balancing Conflicting Values

When city planners try to come up with a transportation plan for a metropolis, much of their work will entail deciding among conflicting values-based visions of the city – preferably in discussion with the widest possible diversity of inhabitants. Bike lanes would lower the pollution levels and perhaps create a less frenetic city, but would require sacrificing some traffic lanes, thus making rush hour worse for drivers. Creating dedicated bus lanes might shorten travel times for riders, but might increase the travel times of people driving their own cars. Turning some major streets into pedestrian walkways might improve the sociality of the city and increase tourism, but might have an impact on traffic and parking. Creating express traffic lanes might improve commute times but reduce local stores' walk-in business. And so on.[4]

Most of these contradictions cannot be resolved. Trade-offs are required. The planning process requires sometimes difficult discussions about the values-based visions of the city. This is an important and not uncommon type of policy-making process.

AI designers engage in similar discussions. For example, the designers of autonomous vehicle systems might put on a white board a list of the values they want their AI-driven machines to support. Top of the list would of course be lowering the number of fatalities. Then, perhaps: shortening travel times, lowering environmental impact, and maintaining comfort. But these may well be incompatible goals: To lower fatalities we might want to lower the vehicles' speed, but that will clearly increase travel times. Drastically lowering the speed might minimize AVs' environmental impact, but fewer people might use them, thus preventing the environmental savings that widespread usage could bring.

Now, the previous paragraph began by saying that the designers of AVs have to decide on the web of values that the system will be trained to support. But, with ML applications such as AVs that are likely to have effects that significantly affect public interests, regulators well may not want to leave critical decisions solely to the designers of these vehicles. Certainly, the Participatory Design approach – which the Responsible by Design process assumes – includes listening to the diversity of voices from those who will be affected by the decision.

The Responsible by Design process deals with values in many of the same ways as policy makers[5], and can benefit from not only the knowledge held by policy makers but also their skills. That is an important point of alignment.

④ For an interesting discussion of the risks of bias in job application systems, see Ifeoma Ajun wa, 'Beware of Automated Hiring,' New York Times, Oct. 8, 2019 bit.ly/ajunwaHiring

⑤ For a thoughtful discussion of values-based morality, see Shannon Vallor, Technology and the Virtues (NYC: Oxford University Press, 2016).

ML Doesn't Think Like Us

It is a strength of ML that it can find more correlations than the human brain can handle. It can find correlations not just among two data points but among webs of data points that individually may have no significant effects. ML does not need to start with general rules, and it may not conclude by generating any. The complexity of these systems means that at least for now they may come up with useful, reliable results in ways that we cannot understand.[6]

This can be a challenge for some ML applications. Being Responsible by Design means making informed, explicit decisions about how explicable a system needs to be, including what elements need explanation, how detailed the explanations need to be, and whether there are tools other than explanations that can accomplish the same purposes.

Explanations, after all, are tools. We usually use them to help us solve problems: the nail explains the flat tire because that explanation tells us what we need to do to fix the flat. When explanations are not available, other tools may suffice. For example, inputting a mortgage application with the gender information changed can help determine if that application was rejected because of the original gender data, but without affording a full explanation of how the system came to its conclusion.[7]

⑥ For a light-hearted look at how differently ML *thinks* than we do, see my 'AI doesn't think like us', bit.ly/PAIRthinklike

⑦ See Brett Mittelstadt and Sandra Wachter, 'Could Counterfactuals Explain Algorithmic Decisions Without Opening the Black Box?', Oxford Internet Institute, Jan. 15, 2018 perma.cc/LT4C-AVTW. Also Finale Doshi-Velez, Mason Kortz et al., 'Accountability of AI Under the Law: The Role of Explanation', Nov. 3, 2017 arxiv.org/abs/1711.01134

ML is a Complex System Deployed in Complex Systems

In the world of classical physics, one billiard ball hits another, transferring energy and causing movement with a force and direction that is perfectly predicted by the known and simple laws of motion.

In a complex, dynamic system, a butterfly landing on a flower in Brazil causes a tornado in Texas because the butterfly's effects may increase in power and area by liberating forces dormant in the system. The tornado is neither predictable nor completely traceable back to the flutter of the butterfly's wings.

Machine learning models are complex and sometimes dynamic. They are often used within wildly complex, dynamic systems such as cities, supply chains, biological systems, weather systems, and the like. This means that the outcomes of the AI can have effects that ripple through a system, spreading wide, and gathering steam as they go, unlike more typical causes that lose energy as their effects roll on. For example, a machine learning system might suggest a small change to bus routes that has results that cascade throughout the city's transportation system and then reach out into housing patterns, education, economics, and more.

This is another reason why the Responsible by Design approach recommends the early and deep involvement of a wide diversity of people – central to Participatory Design – for the complexity of social systems comes not from the number of people involved, but from the differences in their outlooks and interests.

For AI design teams, this means that it is difficult to be confident about how people will use a system without working with people whose outlooks and interests may vary widely. As PAIR's People + AI Guidebook makes clear, this holds not only for how people interact with the system, but for decisions about what the system's purpose is and what the right trade offs are.

* * *

For virtually any machine learning project, the distinctive nature of machine learning raises questions of value and policy that should be considered from the start of a project to its end, and often continuously after the system has been deployed.

The Responsible by Design process therefore entails involving not only a wide diversity of people who may be affected by a machine learning system – as per Participatory Design – but also those who can bring their policy-based skills and expertise to bear on what may be difficult questions of interests, effects, and values.

David Weinberger, Ph,.D., is a writer-in-residence at Google PAIR. Over the past twenty years he has written a series of influential books about the effect of technology on ideas. His latest book, *Everyday Chaos* (2019), looks at how machine learning models may be altering our ideas about how the future happens. He has a Ph.D. in philosophy from the University of Toronto.

His opinions are his own and do not necessarily reflect those of Google.

Notes

Notes

Colophon

Produced as part of People + AI Research Symposium 2019
Held in London on 14 November 2019

Printed in London at Hato Press, using an environmentally friendly stencil, Risograph printer on environmentally sustainable papers and rice based inks

Design by HATO
Illustrations by Alec Doherty