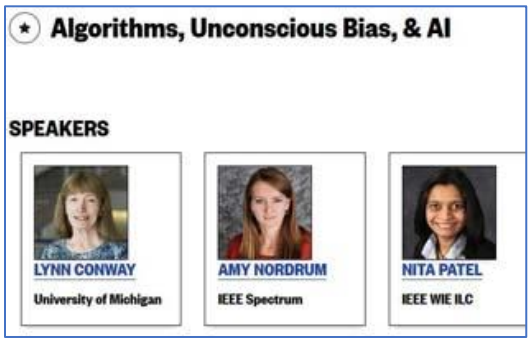
#IEEEAIBias: Algorithms, Unconscious Bias & AI

Panel at SXSW2018: Lynn Conway's Notes



Intro (Amy): (2 mins)

Intro (Amy): (2 mins)

Hi everyone and welcome. I'm Amy Nordrum, news editor at IEEE Spectrum. We're here to talk and learn about how human biases including racism, sexism, and ageism are too often infused into computer models and artificial intelligence. How software programs and search algorithms discriminate against people of color, or push low-income families deeper into poverty.

And—most importantly—how we can avoid perpetuating these human biases through automated systems as we begin to rely more heavily on algorithms to identify problems, make decisions, and roll out solutions for us.

Today I have with me, Nita Patel and Lynn Conway.

Nita is the director of engineering for L-3 Technologies, Warrior Systems, a company in New Hampshire that makes advanced night vision goggles and precision laser-based systems which they sell to the U.S. military.

Before Nita started working on this military technology, she led a major hardware and software upgrade to the National Weather Service's Doppler radar system, which gathers atmospheric data from 178 sites around the world.

Lynn did pioneering work in computer architecture at IBM in the 1960s and in microchip design at Xerox's renowned Palo Alto research lab in the 1970s. She literally "wrote the book" on very large scale integrated chip-design, or VLSI, used by a generation of computer engineers to learn what had been "the black art of microprocessor design."

Working at the Defense Advanced Research Projects Agency in the 1980s, Lynn led a major DOD effort to evolve a technology-base for modern intelligent weapons systems. She then joined the University of Michigan, where she's now Professor Emerita of Electrical Engineering and Computer Science.

I do want to say that there are a number of excellent books that have come out on the topic of bias in algorithms, and I'd encourage you to read them. Much of what we'll talk about here is discussed in much more detail in Weapons of Math Destruction by Cathy O'Neill, Automating Inequality by Virginia Eubanks, Technically Wrong by Sara Wachter-Boettcher, and Algorithms of Oppression by Safiya Umoja Noble. (Interestingly, none of those authors are white men.)

So, to begin….

Part I: Background & Description of the Problem (8 mins)

Q1 (to Nita): How is this different from the algorithms that have long targeted us with online ads, or the ones that turn up search or social media results that it thinks are best suited for each one of us?

Q2 (to Lynn): So what's wrong with trying to automate certain decisions or processes? We all know how biased humans are, couldn't algorithms strip human biases out of important decisions? And how do you determine if an algorithm is perpetuating or removing bias?

At 80 years old old, I'm the 'old-timer' here … Having lived through enormous technological and social changes … I'll try to provide some longer-term perspectives as we go along ...

Regarding WHAT TO AUTOMATE: Let's use automotive tech as an historical example. Think how auto safety features evolved over the years in response to car users getting killed in crashes.

Initially, humans did 'most of the driving' like they'd done with horsedrawn carriages. However, as engineers figured out ways for cars to do things better and more safely (seat belts, airbags, power steering, now automatic braking systems) those things were gradually embedded in the machines … involving LOTS OF transparently-visible usages over time.

In such vast 'techno-social' evolutionary processes, we must avoid naively automating too much right at the start! JUST SO we must avoid "OVER-ALGORITHMITIZING" to quickly, and instead interactively and transparently EMBED AI-things as we go along.

ONE BIG PROBLEM: When unexpected biases get into an AI-augmented product, they can make it unpredictable and can constrain or even hurt users, especially when it's 'out of sight' and INVISIBLE to those users .... who only notice the weirdness when something bizarre happens or goes terribly wrong.

Q3 (to both): What's an example of a system you've read or hear about that you think has crossed the line? What could have been done to prevent or catch it before it was released?

[Lynn goes 2nd]

The "FACEBOOK FIASCO" is a classic … began when Facebook launched its algorithm-based "news feed" to send 'news of interest' to individual users. Bad actors quickly exploited it, and huge waves of "fake-news" began swamping users' pages.

So, what did Facebook do? They used algorithms to search for and stamp out the "fake news"! It was like watching a cat chase its own tail, pouring gasoline on the fire each time around.

In any new technology, we learn through our mistakes. We study what went wrong, learn from it, and develop new methods to prevent it. Hopefully, there won't be many AI algorithmic "Titanics".

PROBLEM: tech folks who trigger a big problem often try to fix it using ONLY what they already know INSIDE THE TECH SILO… as when Facebook's used algorithms to fix a mess caused by algorithms… rather than than thinking outside their TECH-BOX

Q4 (to both): How should we determine what tasks and decisions are automated and which are left to humans?

[Lynn goes 2nd]

Big lessons can be learned from the 'UX' or USER-EXPERIENCE movement now sweeping through the product design community. Instead of designing just 'technological things', UX design

teams ALSO map out and design the ‘user experience’ of using those things . . . involving expanding sets of newbie-trial-users during the design process.

UX design teams quickly discover which subtasks are best automated and which are best done by humans. The goal of ‘automation’ is not to remove the human from the process, but to empower humans who use and build on that ‘automation’!

These methods ALSO HELP AVOID leaving the human usage decisions up to the algorithm coders who implement them. The UX design team takes ‘ownership’ of a products TECHNO+SOCIAL behavior, not the coders.

Part II: Why and How Algorithms Become Biased

Q5 (to Nita): What are some of the common mistakes that programmers make that cause these algorithms to become biased?

Q6 (to Lynn): Once you’ve built a predictive model or written an algorithm to do something useful, how do you tell whether it’s evil?

Designing an algorithm is like creating a recipe, writing an instruction manual or preparing a talk. You better have your audience in mind, and the more you test it in early trials, the better. Once launched, you must react quickly to user feedback if things ‘break’ or produce harmful results.

BUT SOMETIMES CAN BE DIFFICULT: For example, imagine you’re giving a carefully prepared talk in a foreign country … accidentally launch the ‘wrong words’ into that audience … and totally out of the blue accidentally trigger a huge riot!

So,a recipe, an algorithm is just a ‘thing’. In itself it’s neither useful nor evil. It only becomes ‘animated’ and ‘observable in action’ when it’s running and entangled in a particular context. And its effects might be totally different in different contexts.

Q7 (to Nita): Are there certain types of models or algorithms or programs that are more prone to this? Or are there certain types of problems or questions that modelers try to answer that tend to produce biased programs?

Q8 (to Nita, as follow up): A lot of these issues seem to stem from programmers and modelers who assign proxies and how those proxies are used. Why are proxies so problematic, and what causes us to reach for them?

Q9 (to Lynn): In the age of big data, I think we like to believe we can throw math and computer science at anything and improve it. But a key point, as Cathy O’Neill points out in her book, is trying to figure out what you’re trying to model in the first place and then having the right data to speak to that question. How do we know whether we have the right data, or whether data can be meaningfully applied to a problem we’re trying to solve?

In ”GEEK HERESY: Rescuing Social Change from the Cult of Technology”, Kentaro Toyama reveals the frequent blindness of tech-insiders to external social contexts and processes …

Consider too, the “AI Replication Crisis” …. Many methods used to build and apply AI algorithms, especially in “machine-learning”, are based on research that’s now proving to be “non-replicable” … many of these methods are proving to be “brittle” and subject to unexpected failures … even to the point of reports of AI vision systems suffering from sudden “Hallucinations” … i.e., seeing things that aren’t there.

SO: how can we evolve and apply AI System Design Methodologies that include wider and deeper abstractions of AI user-experiences, including users’ entanglement in larger techno + social contexts? And how do we quickly update installed system to account for failure reports?....

Q10 (to Lynn): Whose responsibility is this? Do we get to all just blame the big tech companies, or is it our fault for sharing our data so freely, or is this the sort of thing you can regulate or build standards around?

Visualizing tech as BIG INFRASTRUCTURE vs many diverse COMMERCIAL PRODUCTS …

[i] The INFRASTRUCTURE will follow the historical paths of regulated public transportation and utilities industries … this 'splitting process' will be politicized, painful, but inevitable.

[ii] The enormous future marketplace of AI-enhanced COMMERCIAL PRODUCTS must necessarily include vast user-community engagement and feedback ... bringing ever more people into the development cycle loops of UX product design teams.

IN BOTH CASES new standards must be rapidly evolved to insure utility, safety, efficiency and affordability and avoid monopolistic practices and harmful, dangerous products. Think of it as a whole new world of digital "building codes" that enable us to build good highways and homes, while remaining well connected to the world and also good neighbors.

The IEEE has launched a Global Initiative on Ethics of Autonomous and Intelligent Systems as an incubation space for new standards and solutions, certifications and codes of conduct, and consensus building for ethical implementation of intelligent technologies … something to seriously watch and get involved in if you can.

Q11 (to both): We've seen a lot of examples covered in the media of recidivism software that's racist, job search referral programs that are sexist, and algorithms that target low-income people for predatory loans or for-profit universities. Do you have experience from your own day to day work or careers as engineers that you feel are instructive to others?

[Lynn goes 2nd]

Lots of harmful bias gets embedded in AI as a result of automatic data-gathering by commonly-asking apparently-innocuous questions. Using excess data during training even though that data is irrelevant to the task at hand . . . can introduce 'bias'.

For example, we check the box 'MALE' or 'FEMALE' on every imaginable bureaucratic form we ever fill out. That 'data' is then entered into all sorts of software records, studies, sortings and decisions … that are now being revealed as gender-biased.

SO: Why is it that we're asked to enter our gender, race, age, income, etc., even so many forms where that data's totally irrelevant to the task at hand? . . . NOW, think about all the data advertisers have about you! And how that accumulating PROFILE might bias decisions about you in all sorts of UNANTICIPATED ways!

Q12 (to Lynn): How much of this is caused by the demographics of the people coding—would we have the same problems if we had more minorities building models?

BIG QUESTION: Just who are the people training all this AI stuff? … Do they have degrees in computer-science, expertise in 'deep-learning' models and algorithms? NOPE, not the case . . .

Tons of what appears to be AI right now was done by humans using Amazon's 'Mechanical Turk' system ... as a kind of "Artificial Artificial Intelligence" where humans behind the scenes do tasks computers don't do well....

Meantime, Google just announced the Google ML machine-learning cloud service enabling 'developers' with limited AI experience to train "high-quality AI models" as more of tasks become automatable.

This reveals a BIG HIDDEN DANGER: by leaving key social-process decisions up to anonymous algorithm trainers and coders . . .we have no clue who's LAUNCHED that stuff into the world … Thus NO ONE CAN BE HELD ACCOUNTABLE when things go terribly wrong, AND WE HAVE FEW CLUES about what biases have crept into that AI.

One solution is far wider use of UX design teams that design such integrated technological + social processes, that then monitor the implementation of those products by algorithm trainers and coders, and take responsibility' for the techno-social products' behaviors.

Just think of how big bridges are built: their users are seriously taken into account every step of the way,, and the public can figure out who's responsible if they ever collapse..

Part III: What's to be Done?

Q13 (to Nita): So what steps can someone who's working with data take to make sure they don't build one of these?

Q14 (to both) What questions should we ask when we hear something is being automated or running on AI or machine learning? How should we examine and stress test these systems?

[Lynn goes 2nd]

Looking at the larger picture: How can we check out any new complex system:

(i) check into "WHO MADE IT" and what their reputation is, (ii) find out whether and how THEY benchmarked and "STRESSED-TESTED IT",

(iii) compare THEIR reputation and product development methods with their COMPETITORS in the same niche, (iv) learn about emerging problems in THAT PRODUCT-NICHE and whether that product may be having them, (v) think about YOUR possible EXPOSURES AND LIABILITIES if their product 'misbehaves' . . .

It's much like checking the reviews of a new cars' features before "buying it", whether at an individual, community, or corporate level …

Q15 (to both): And for someone who's not working as a data scientist or a modeler, what can they do?

[Lynn goes 2nd]

Think ahead to where this is headed . . . Remember how what appeared to be AI machine learning was actually "Artificial Artificial Intelligence" where human Mechanical Turks did things computers couldn't do at the time? ...

THINK ABOUT Google's new machine-learning service, enabling users to build and train their own customized AI 'deep learning' models … so that more things can be routinely done by AI....

Eventually many people will routinely train modest-scale every-day AI-animated things to support and empower themselves in their personal, family and community lives. They'll be the "recipe-makers" of the coming era. The coolest 'recipes' will go viral, evolving as they spread.

As a result, we're at the onset of a new techno-social age . . . where ALL OF US can participate in this accelerating evolutionary process . . . if we're able to 'JUMP ON THAT WAVE'...

So, we all have some serious choices: We can EITHER "SHELTER FROM THE COMING HIGH-TECH STORMS or GET OUT ON FRONT OF AND SURF THEIR BIG WAVES!!

The bottom line [image@ #IEEEAIBias]