As over-hyped as artificial intelligence is—everyone's talking about it, few fully understand it, it might leave us all unemployed but also solve all the world's problems—its list of accomplishments is growing. AI can now write realistic-sounding text, give debating champs a run for their money, diagnose illnesses, and generate fake human faces—among much more.

After training these systems on massive datasets, their creators essentially just let them do their thing to arrive at certain conclusions or outcomes. The problem is that more often than not, even the creators don't know exactly *why* they've arrived at those conclusions or outcomes. There's no easy way to trace a machine learning system's rationale, so to speak. The further we let AI go down this opaque path, the more likely we are to end up somewhere we don't want to be—and may not be able to come back from.

In a panel at the South by Southwest interactive festival last week titled "Ethics and AI: How to plan for the unpredictable," experts in the field shared their thoughts on building more transparent, explainable, and accountable AI systems.

Not New, but Different

Ryan Welsh, founder and director of explainable AI startup Kyndi, pointed out that having knowledge-based systems perform advanced tasks isn't new; he cited logistical, scheduling, and tax software as examples. What's new is the learning component, our inability to trace how that learning occurs, and the ethical implications that could result.

"Now we have these systems that are learning from data, and we're trying to understand why they're arriving at certain outcomes," Welsh said. "We've never actually had this broad society discussion about ethics in those scenarios."

Rather than continuing to build AIs with opaque inner workings, engineers must start focusing on explainability, which Welsh broke down into three subcategories. Transparency and interpretability come first, and refer to being able to find the units of high influence in a machine learning network, as well as the weights of those units and how they map to specific data and outputs.

Then there's provenance: knowing where something comes from. In an ideal scenario, for

example, Open AI's new text generator would be able to generate citations in its text that reference academic (and human-created) papers or studies.

Explainability itself is the highest and final bar and refers to a system's ability to explain itself in natural language to the average user by being able to say, "I generated this output because x, y, z."

"Humans are unique in our ability and our desire to ask why," said Josh Marcuse, executive director of the Defense Innovation Board, which advises Department of Defense senior leaders on innovation. "The reason we want explanations from people is so we can understand their belief system and see if we agree with it and want to continue to work with them."

Similarly, we need to have the ability to interrogate AIs.

Two Types of Thinking

Welsh explained that one big barrier standing in the way of explainability is the tension between the deep learning community and the symbolic AI community, which see themselves as two different paradigms and historically haven't collaborated much.

Symbolic or classical AI focuses on concepts and rules, while deep learning is centered around perceptions. In human thought this is the difference between, for example, deciding to pass a soccer ball to a teammate who is open (you make the decision because conceptually you know that only open players can receive passes), and registering that the ball is at your feet when someone else passes it to you (you're taking in information without making a decision about it).

"Symbolic AI has abstractions and representation based on logic that's more humanly comprehensible," Welsh said. To truly mimic human thinking, AI needs to be able to both perceive information and conceptualize it. An example of perception (deep learning) in an AI is recognizing numbers within an image, while conceptualization (symbolic learning) would give those numbers a hierarchical order and extract rules from the hierachy (4 is greater than 3, and 5 is greater than 4, therefore 5 is also greater than 3).

Explainability comes in when the system can say, "I saw a, b, and c, and based on that

decided x, y, or z." DeepMind and others have recently published papers emphasizing the need to fuse the two paradigms together.

Implications Across Industries

One of the most prominent fields where AI ethics will come into play, and where the transparency and accountability of AI systems will be crucial, is defense. Marcuse said, "We're accountable beings, and we're responsible for the choices we make. Bringing in tech or AI to a battlefield doesn't strip away that meaning and accountability."

In fact, he added, rather than worrying about how AI might degrade human values, people should be asking how the tech could be used to help us make better moral choices.

It's also important not to conflate AI with autonomy—a worst-case scenario that springs to mind is an intelligent destructive machine on a rampage. But in fact, Marcuse said, in the defense space, "We have autonomous systems today that don't rely on AI, and most of the AI systems we're contemplating won't be autonomous."

The US Department of Defense released its 2018 artificial intelligence strategy last month. It includes developing a robust and transparent set of principles for defense AI, investing in research and development for AI that's reliable and secure, continuing to fund research in explainability, advocating for a global set of military AI guidelines, and finding ways to use AI to reduce the risk of civilian casualties and other collateral damage.

Though these were designed with defense-specific aims in mind, Marcuse said, their implications extend across industries. "The defense community thinks of their problems as being unique, that no one deals with the stakes and complexity we deal with. That's just wrong," he said. Making high-stakes decisions with technology is widespread; safety-critical systems are key to aviation, medicine, and self-driving cars, to name a few.

Marcuse believes the Department of Defense can invest in AI safety in a way that has far-reaching benefits. "We all depend on technology to keep us alive and safe, and no one wants machines to harm us," he said.

A Creation Superior to Its Creator

That said, we've come to expect technology to meet our needs in just the way we want, all the time—servers must never be down, GPS had better not take us on a longer route, Google must always produce the answer we're looking for.

With AI, though, our expectations of perfection may be less reasonable.

"Right now we're holding machines to superhuman standards," Marcuse said. "We expect them to be perfect and infallible." Take self-driving cars. They're conceived of, built by, and programmed by people, and people as a whole generally aren't great drivers—just look at traffic accident death rates to confirm that. But the few times self-driving cars have had fatal accidents, there's been an ensuing uproar and backlash against the industry, as well as talk of implementing more restrictive regulations.

This can be extrapolated to ethics more generally. We as humans have the ability to explain our decisions, but many of us aren't very good at doing so. As Marcuse put it, "People are emotional, they confabulate, they lie, they're full of unconscious motivations. They don't pass the explainability test."

Why, then, should explainability be the standard for AI?

Even if humans aren't good at explaining our choices, at least we can try, and we can answer questions that probe at our decision-making process. A deep learning system can't do this yet, so working towards being able to identify which input data the systems are triggering on to make decisions—even if the decisions and the process aren't perfect—is the direction we need to head.

Image Credit: a-image / Shutterstock.com

By Vanessa Bates Ramirez

This article originally appeared on Singularity Hub, a publication of Singularity University.