Bias on Both Sides: Decision Making with an AI Advisor

Dr. Joel Davis

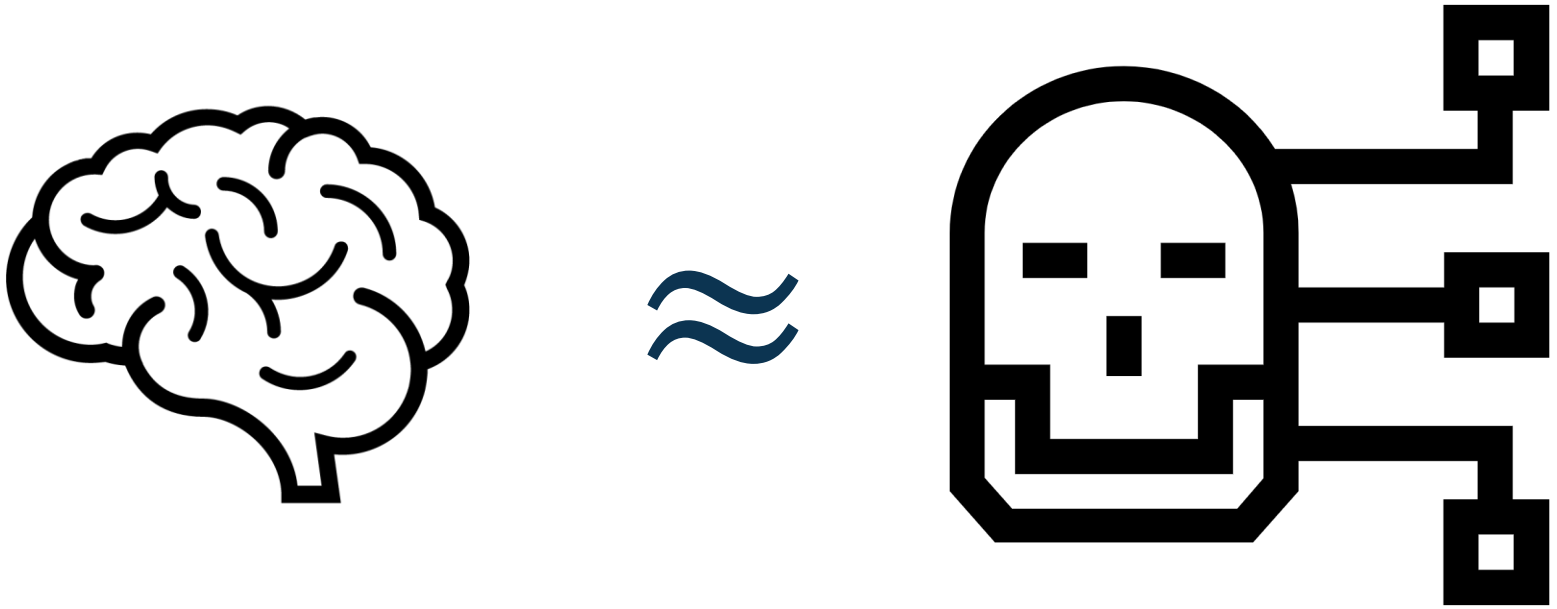
University of Florida

Joel.davis@warrington.ufl.edu

Abstract

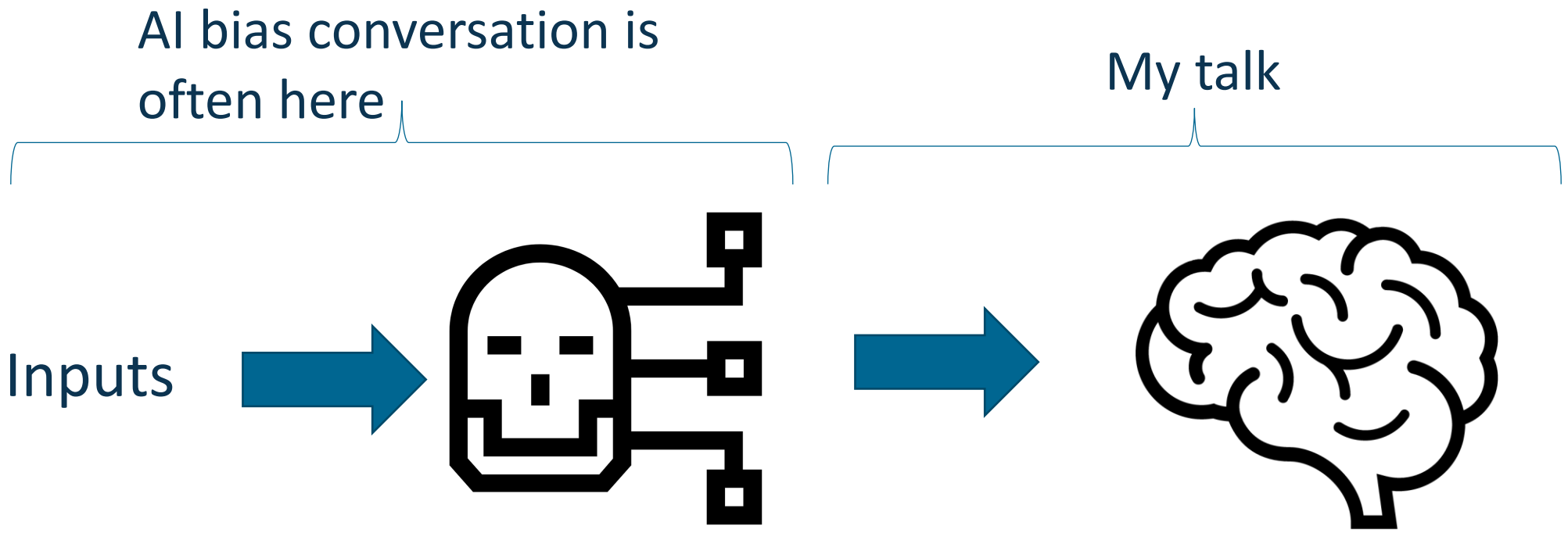
While most of the attention in AI bias focuses on the algorithm and the data, this discussion will explore the bias humans have against and sometimes for using AI and algorithmic systems designed to assist in making decisions. These AI advisors can be found in many domains. For example, a semi-autonomous financial advisor that generates and delivers a recommendation to a human, who is ultimately responsible for judging that advice and performing an action. Some have suggested that humans exhibit algorithm aversion and therefore avoid using algorithms to help make decisions, even when the algorithm is verifiably better. The discussion will focus on preliminary research to understand decision-makers' complex and sometimes conflicting behavior when it comes to using the advice generated by AI.

One way of thinking about AI is that it is an attempt to replicate human intelligence.



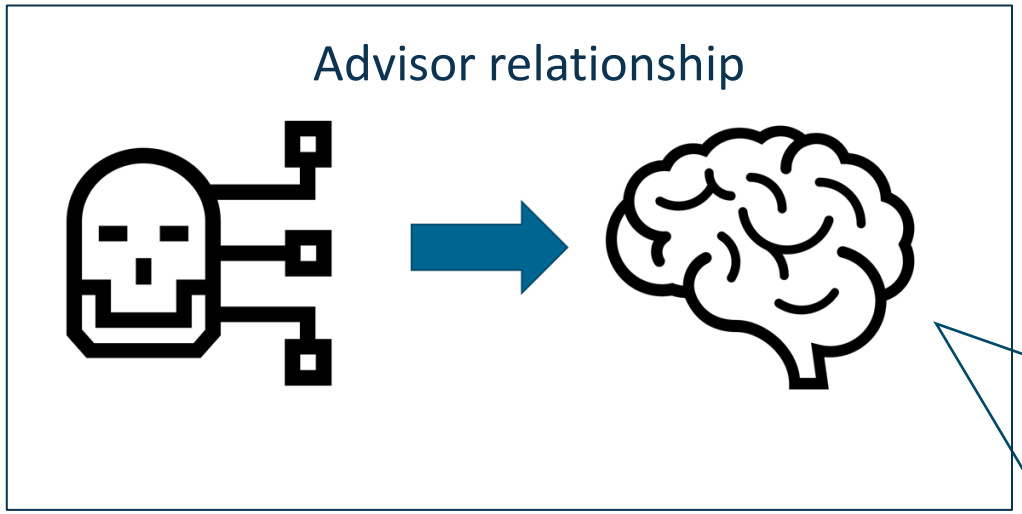
Once we've done that… we send these AI agents out into the world, and they operate independently.

Another way to view this is that most AI operate *with* humans, and often are used to help support humans in decision making.



Why is this important?

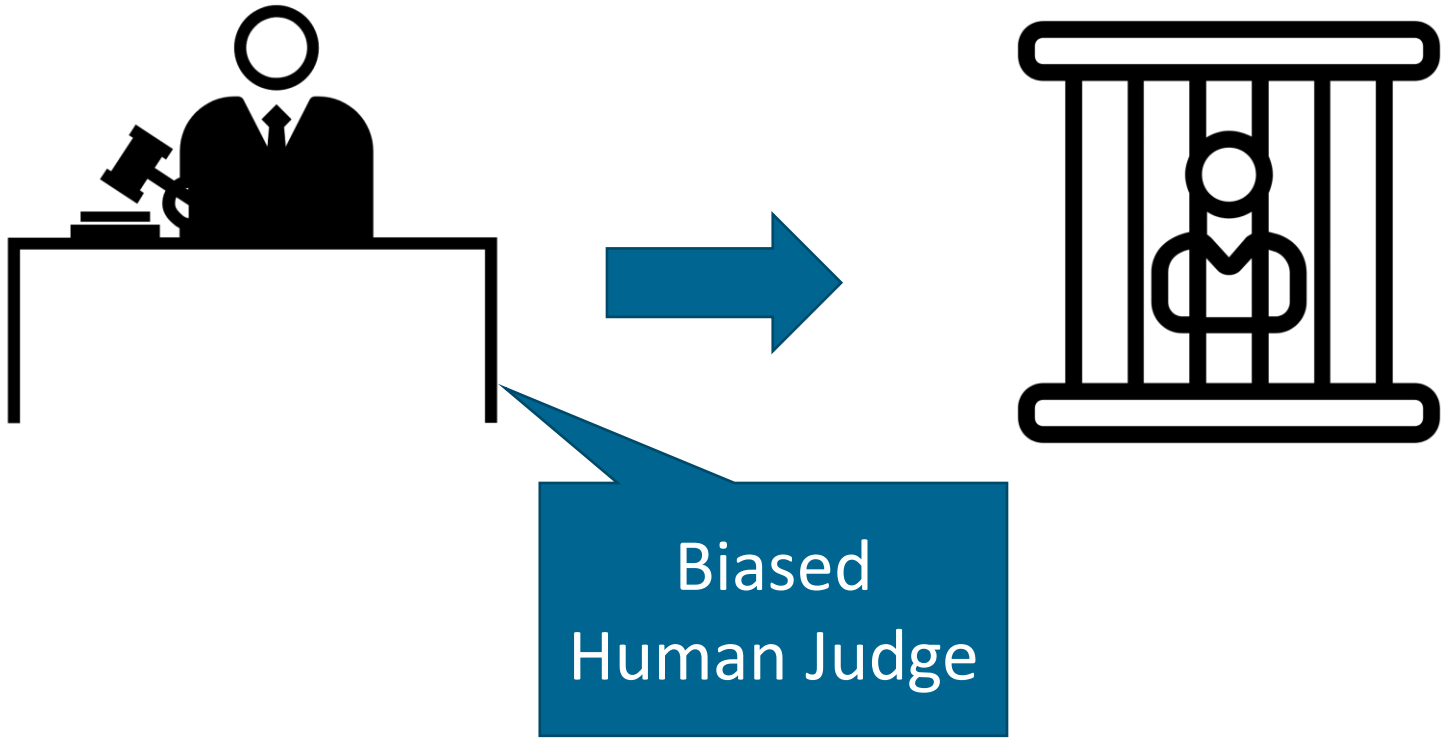
- Chatbots (yes)



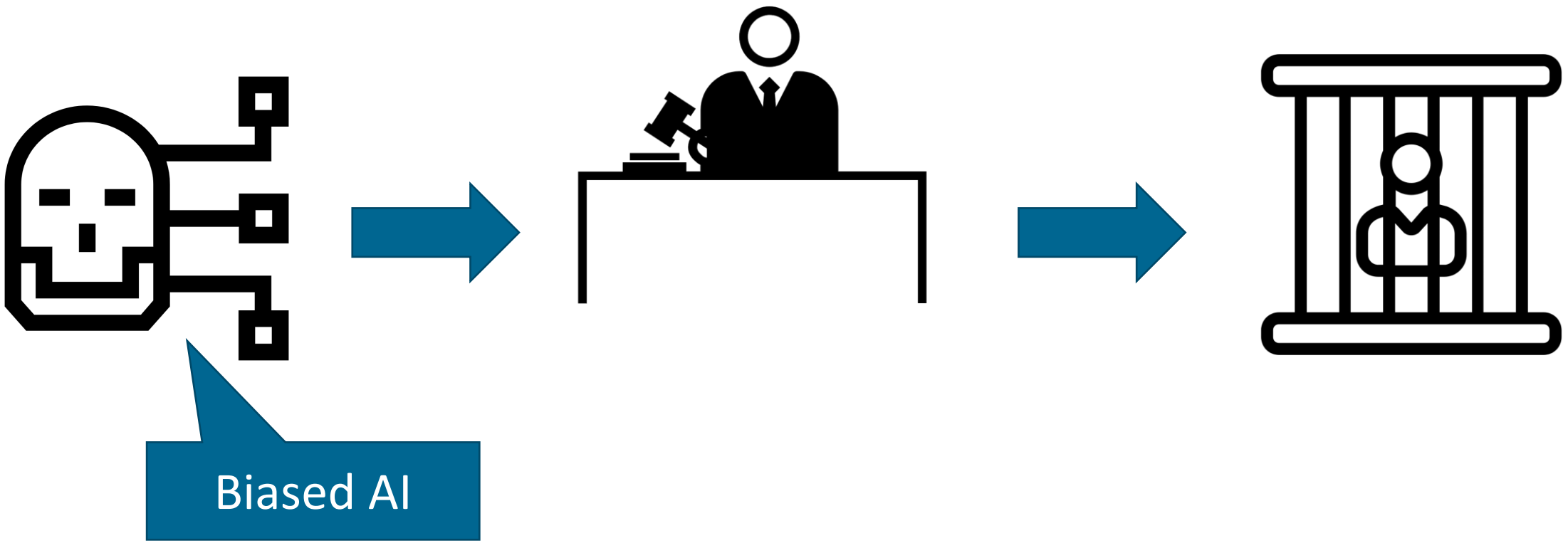
Also…

- Self-driving vehicles
- Recommendation engines
- Legal decision support tools (like estimators of recidivism)
- Business forecasting solutions
- Etc…

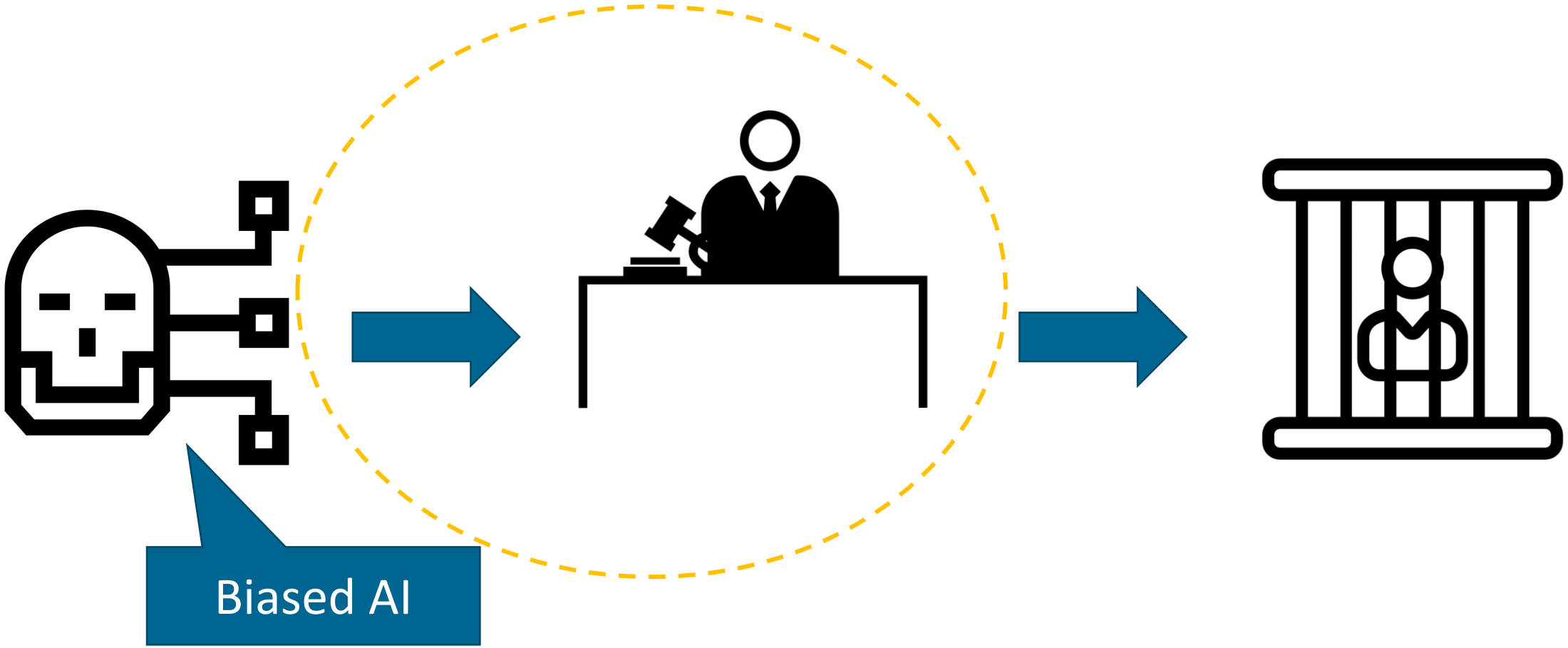
AI advisors, in the courtroom were supposed to reduce or remove human bias from decision making.



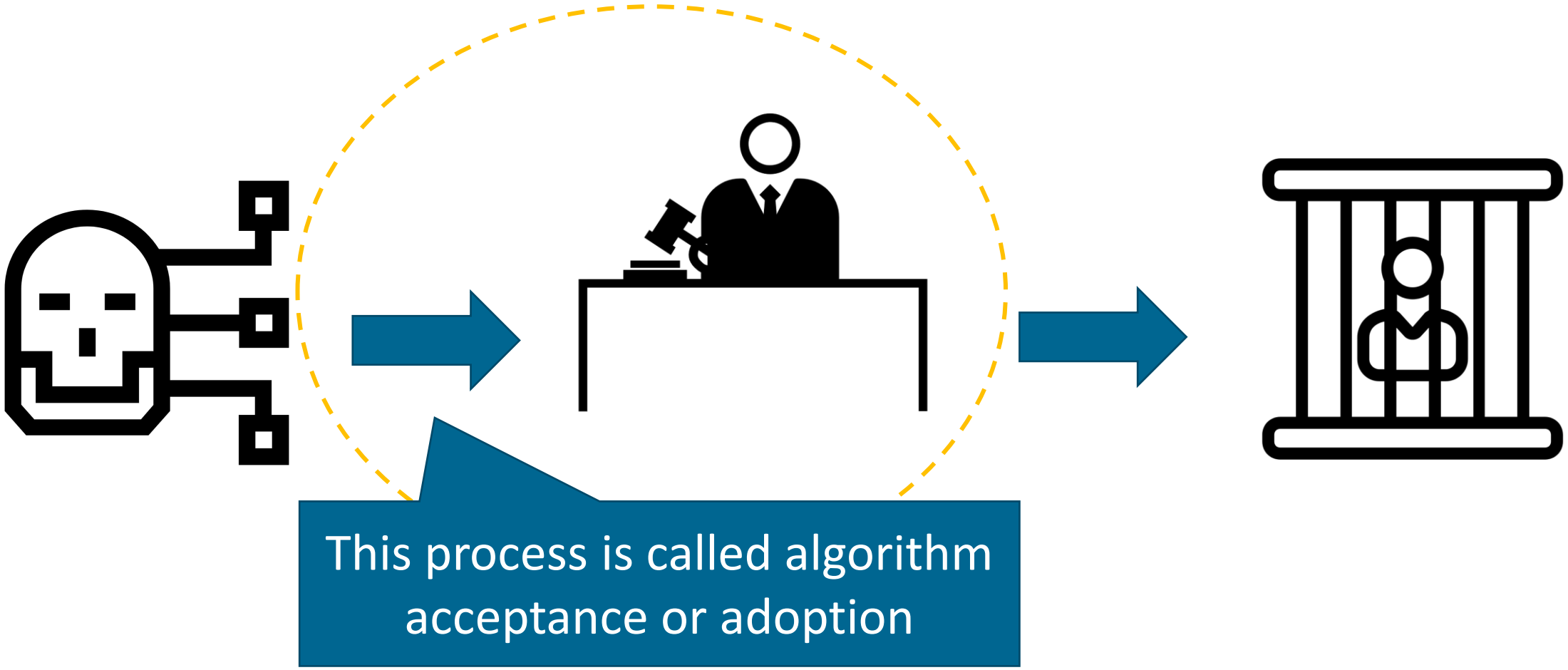
Obviously, that didn't work.



While we pay a lot of attention to the bias coming from the algorithm, we tend to ignore the new bias in the judge.



While we pay a lot of attention to the bias coming from the algorithm, we tend to ignore the new bias in the judge.



Initial results of qualitative study on algorithm/AI adoption by decision makers.

- Interviewed 50 decision makers in business. Their work experience ranged between 1955 and 2020, with a minimum of 2 years and average work experience of just over 17 years.
- All of the subjects had experience using the advice or output from an algorithm and making decisions based on that advice.

Factors in rejecting/accepting the output of an algorithm (50 interviews):

78%

Makes Sense*

The outputs of the algorithm have to **make sense.** I considered this part of factor I called **"Output Trust"**

64%

Input Trust

Evaluated the number or *perceived* quality of the inputs to the algorithm.

68%

Algorithm Provenance

The credibility of the creator, development process, and history of the system.

50%

Understandability

A combination of interpretability* and explainability.

* interpretability + makes sense was explored in great detail in these interviews. But the connection between the 2 was much weaker than I assumed it would be. "Makes Sense" seems to come from prior domain experience or gut, and the more senior and experience people cared less about interpretability, more about it aligning with their preconceived notions. (looks a lot like confirmation bias!)

Factors in rejecting/accepting the output of an algorithm (50 interviews):

78%

Makes Sense*

The outputs of the algorithm have to **make sense.** I considered this part of factor I called **"Output Trust"**

64%

Input Trust

Evaluated the number or *perceived* quality of the inputs to the algorithm.

68%

Algorithm Provenance

The credibility of the creator, development process, and history of the system.

50%

Understandability

A combination of interpretability* and explainability.

* interpretability + makes sense was explored in great detail in these interviews. But the connection between the 2 was much weaker than I assumed it would be. "Makes Sense" seems to come from prior domain experience or gut, and the more senior and experience people cared less about interpretability, more about it aligning with their preconceived notions. (looks a lot like confirmation bias!)

Factors in rejecting/accepting the output of an algorithm (50 interviews):

78%

Makes Sense*

The outputs of the algorithm have to **make sense.** I considered this part of factor I called **"Output Trust"**

64%

Input Trust

Evaluated the number or *perceived* quality of the inputs to the algorithm.

68%

Algorithm Provenance

The credibility of the creator, development process, and history of the system.

50%

Understandability

A combination of interpretability* and explainability.

* interpretability + makes sense was explored in great detail in these interviews. But the connection between the 2 was much weaker than I assumed it would be. "Makes Sense" seems to come from prior domain experience or gut, and the more senior and experience people cared less about interpretability, more about it aligning with their preconceived notions. (looks a lot like confirmation bias!)

Factors in rejecting/accepting the output of an algorithm (50 interviews):

78%

Makes Sense*

The outputs of the algorithm have to **make sense.** I considered this part of factor I called **"Output Trust"**

64%

Input Trust

Evaluated the number or *perceived* quality of the inputs to the algorithm.

68%

Algorithm Provenance

The credibility of the creator, development process, and history of the system.

50%

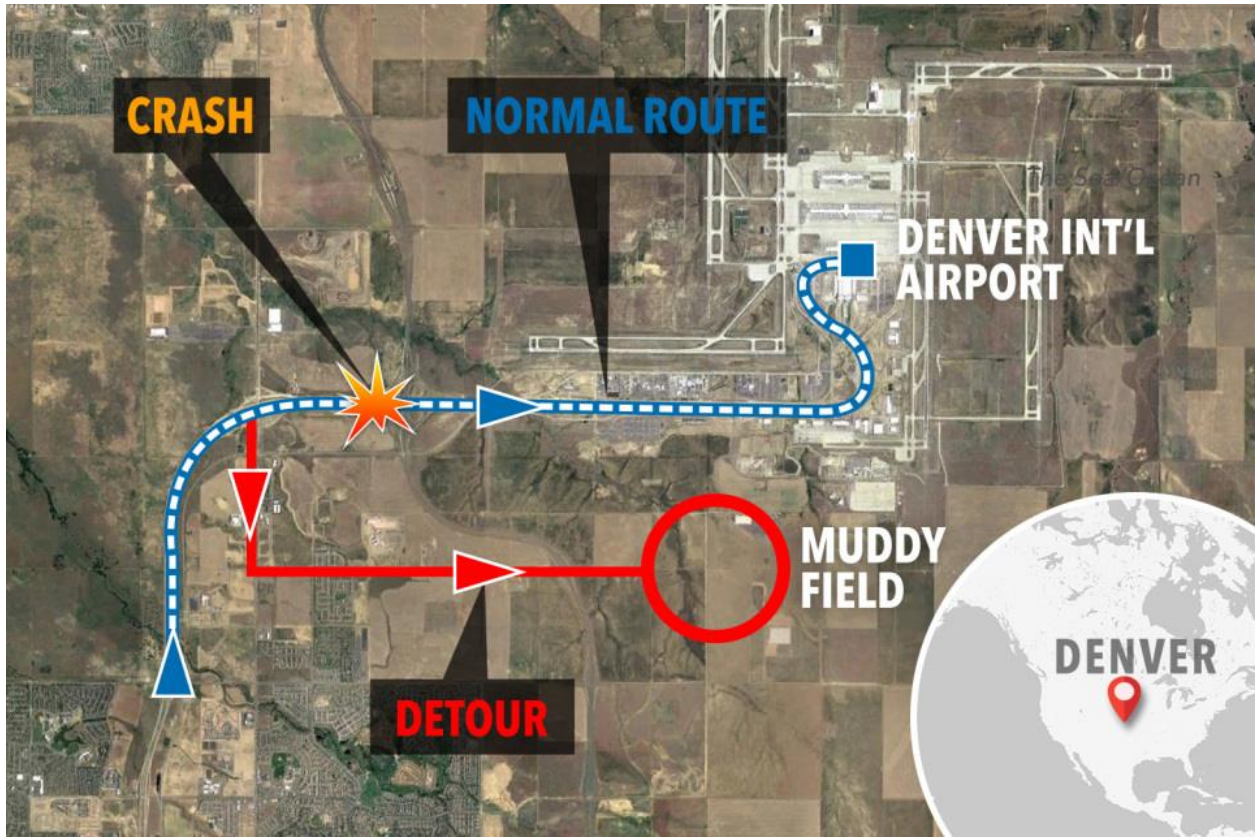
Understandability

A combination of interpretability* and explainability.

* interpretability + makes sense was explored in great detail in these interviews. But the connection between the 2 was much weaker than I assumed it would be. "Makes Sense" seems to come from prior domain experience or gut, and the more senior and experience people cared less about interpretability, more about it aligning with their preconceived notions. (looks a lot like confirmation bias!)

90% said *other* decision makers practice "algorithmic deference", just taking the advice of the machine.

- None of the subjects expressed that **they themselves** ever practiced this sort of deference



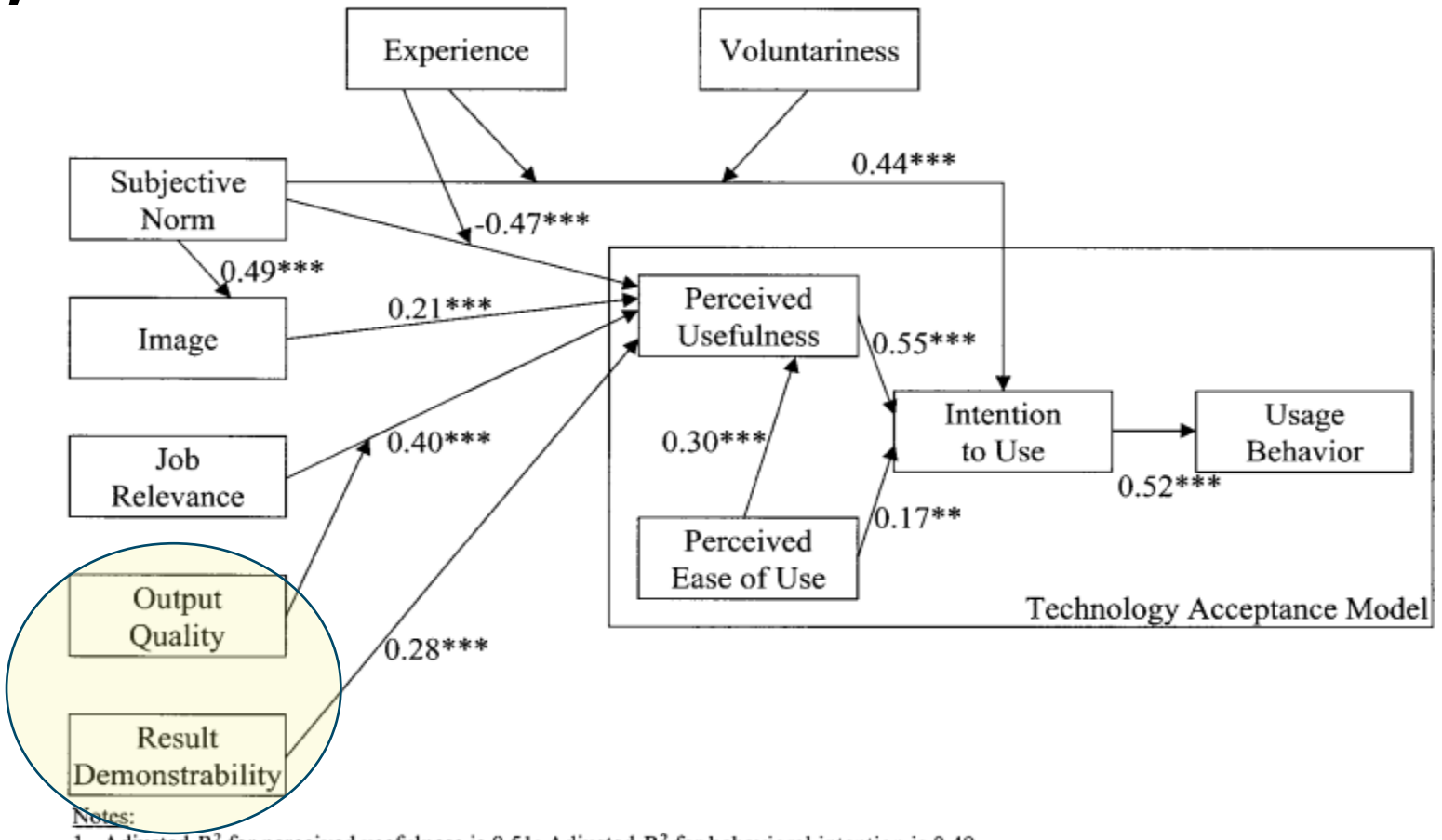
"Dopey drivers rushing to airport blindly follow Google Maps 'shortcut' taking them to muddy FIELD piled with 'a hundred' stranded cars"

https://www.thesun.co.uk/news/9382403/airport-denver-google-maps-shortcut-stranded/

Why shouldn't we just use the Technology Acceptance Model? (TAM)

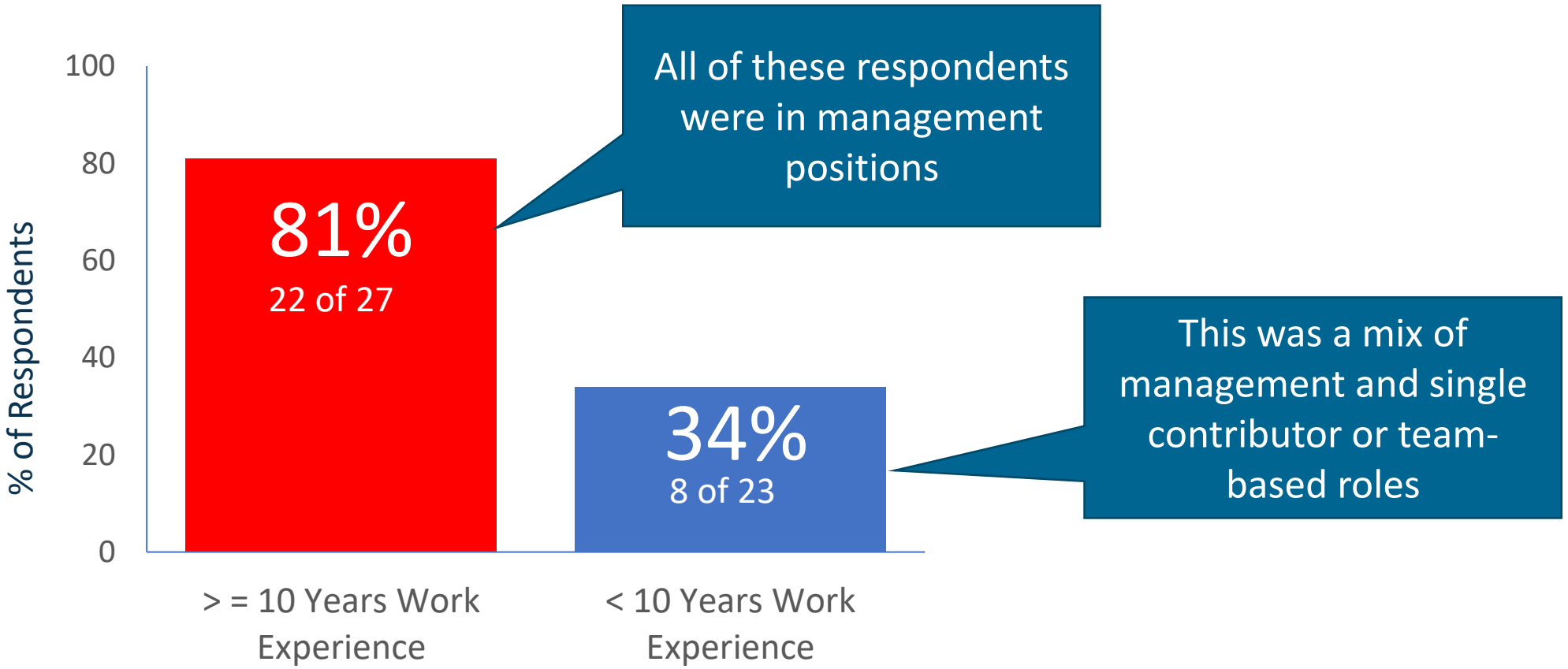
Algorithm Provenance

Makes sense- Output Trust

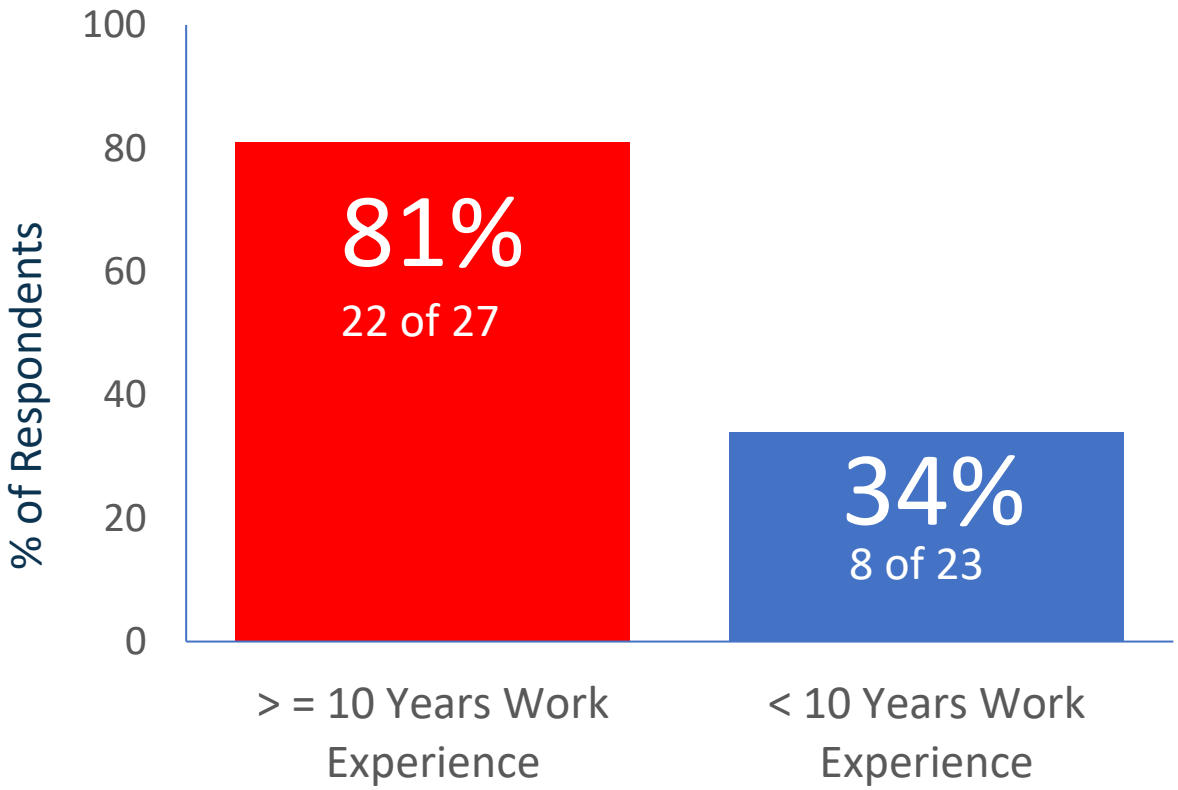


Venkatesh, V., & Davis, F. D. (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. Management Science, 186–204.

Respondents in the interviews that expressed they could not assess the quality of the AI or algorithmic system



Respondents in the interviews that expressed they could not assess the quality of the AI or algorithmic system



"Basically, to me, an algorithm is here to validate mathematically my gut feel, right? My gut feel tells me this right? Now I have all these inputs that are coming from those algorithms. Do those two make sense?"

Makes sense- Output Trust

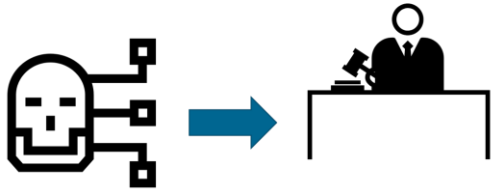
"Because typically with any type of algorithm, you do have some type of expected result. If it's in a range of what you would expect, then I guess that would be an answer of yes, this is accurate"

Makes sense- Output Trust

"You don't need explicit understanding. I would say that when I use something that somebody else created, I assume that they're much smarter than me and they know what they're doing."

Algorithm Provenance

Where should we go from here?



- There is a significant amount of research on understanding and addressing human bias (confirmation, anchoring, availability etc.)
- Clearly education about cognitive biases is critical
- There is very little research on how the *interaction* with a *machine* changes these biases, and how to address that problem.
- We need more conversations/research and effort related to the challenges associated with decision makers using or adopting AI effectively.

Reducing bias in AI requires an approach that INCLUDES how humans receive and use the outputs of the process.

It is not enough to reduce the bias in the data, or in the algorithm process, because many of these systems still have a human in the loop. A human that can and will… re-inject bias.