

Overcoming Automation Bias

A body of learning designed to raise the general awareness about over-reliance of AAA Systems as required by the EU AI Act for human oversight



FORHUMANITY

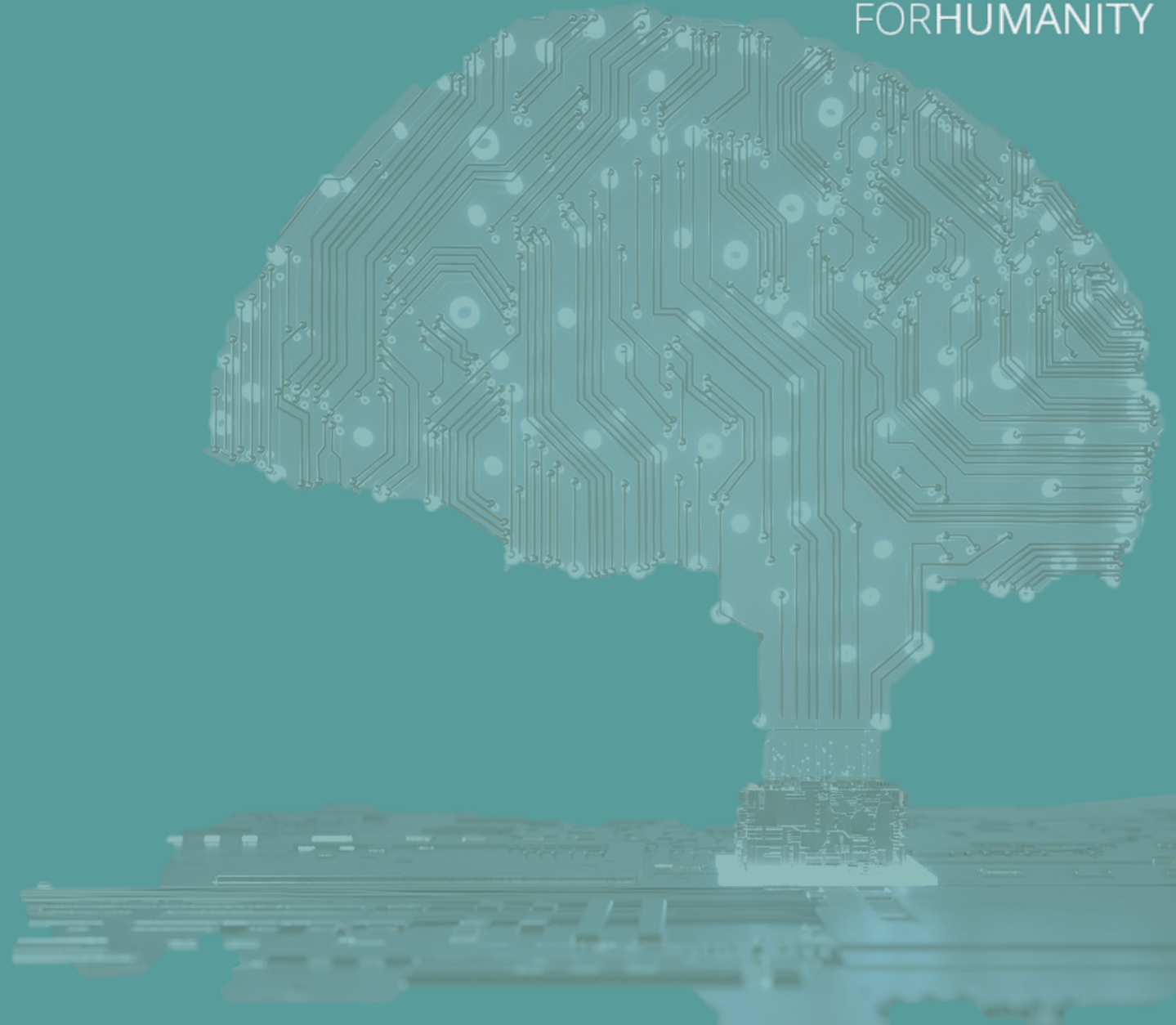




FORHUMANITY

Agenda

1. Introduction to the course
2. What is Automation Bias?
3. TBD . . .



1. Introduction to the Course

The curriculum is designed to establish a healthy skepticism in regards to **AAA Systems** and to educate users when **AAA Systems** can be relied upon (in the form of an infrastructure of trust). This skill is especially important for human oversight involving overriding, stopping or pausing an AAA System

The curriculum further encourages users to acquire knowledge and understanding of underlying assumptions, data inputs, risk mitigations, and residual risk associated with the **AAA System**



FORHUMANITY

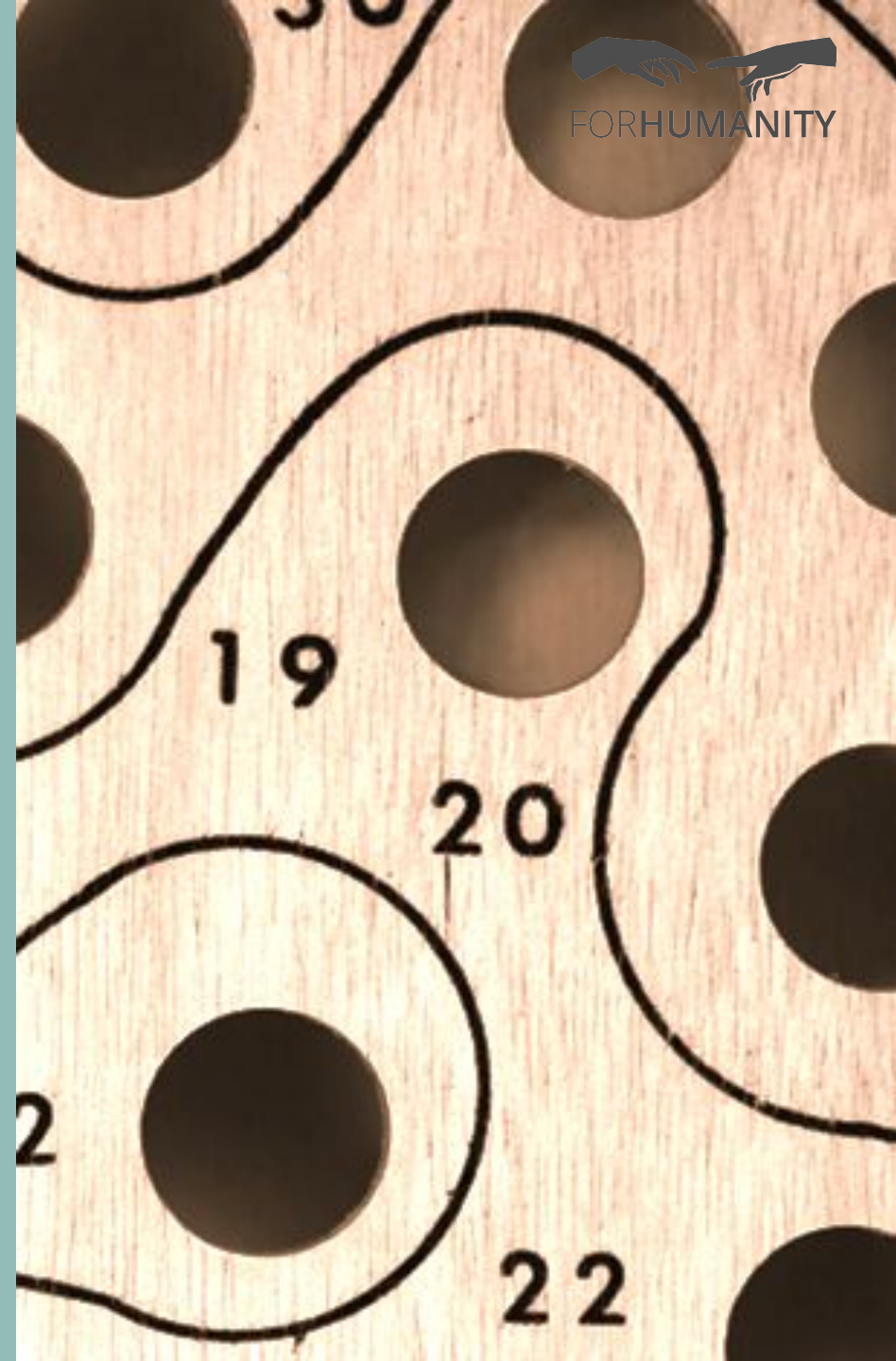


Purposes and Objectives

This presentation is the starting point for your class

The purpose of this presentation is to create a framework for teaching about Automation Bias. It is designed to be a starting point for your teaching slides.

The content that follows represent our collective establishment of learning objectives about what Automation Bias is and associated tools to overcome it. All licensees should feel comfortable to develop their own materials or borrow directly from the information provided here as long as the primary messages (learning objectives) are delivered comprehensively.



Goals of this Course



Awareness

Become aware of the existence of Automation Bias



Understanding

Understand the implications and risk of Automation Bias



Balance

Learn a healthy balance between trust and doubt



Mastery

Gain the skills and tools to overcome Automation Bias



When tools can and should be relied upon

When tools require human oversight

2. What is Automation Bias?

Automation Bias is the propensity for humans to over trust and/or over rely on AAA Systems even when presented with contradictory knowledge and information from other reliable sources such as human senses and/or experience



The History of Automation Bias

The knowledge that the roots of many automation-related errors can be traced to cognitive shortcuts has major implications for training interventions.”

- Linda Skitka

- Term is coined in mid-1990s in a study about aviation (Mosier and Skitka)
- Researchers were interested in the ways humans relied on technology and how that led to errors
- They studied the impacts of aviation technology on pilots
- Humans tend to travel the path of least cognitive resistance
- More pervasive and complex machines = greater propensity for automation bias
- Now we're building machines that are more human-like than ever
- Leading to issues with Anthropomorphism, which both amplifies and obscures Automation Bias

Photo by Clarisse Meyer on Unsplash



Anthropomorphism is Not Inherently Bad

But it can be exploited in dangerous ways

It is a basic human characteristic at a fundamental, biological level, driven by hormone hits in our brain's pleasure centers

But like so many other similar biological level drives, it can be exploited - sometimes for good, but also for selfish reasons

Anthropomorphism makes us susceptible to Automation Bias - We can set ourselves up for that exploitation because we over-trust and we over-rely upon AAA Systems

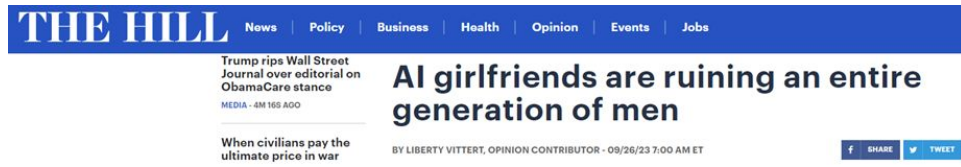


It's Not Our Duty to Stop Anthropomorphism

We cannot prevent people from desiring and designing machines with human characteristics - And we won't try



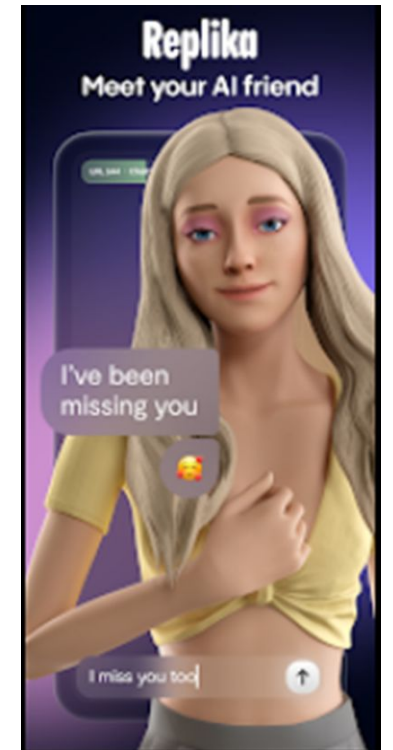
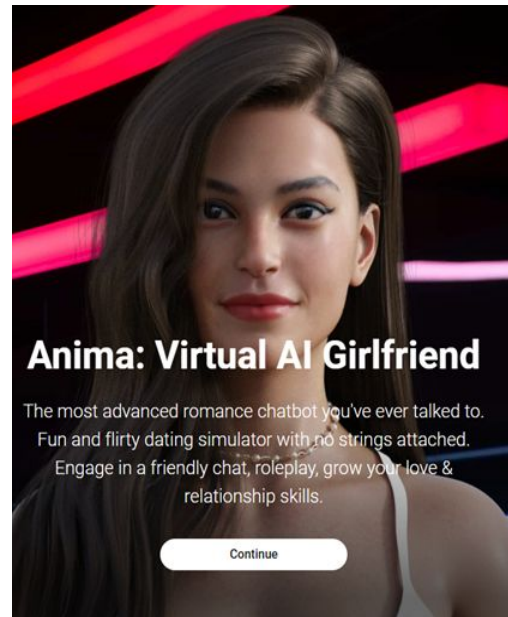
It is our duty to ensure that the risks of such behavior are well-understood and the benefits of other, human-first choices comes with benefits that celebrate our humanity



The surprising case for AI boyfriends

APRIL 4, 2023 · 6:33 PM ET

By Jessica Mendoza, Brittany Luse, Jessica Placzek, Veralyn Williams



How is Automation Bias Dangerous to the Well-being of Humans?



Video being produced - expected ETA March 1

Automation Bias Can Come From Everyone and Everywhere

It can occur in many levels, many places

Automation Bias can manifest through designers, developers and deployers of AAA Systems

But it can also manifest for users - the pilots of Air France flight 447 didn't design, develop, or deploy the AAA System that caused the fatal crash, they were users of the System

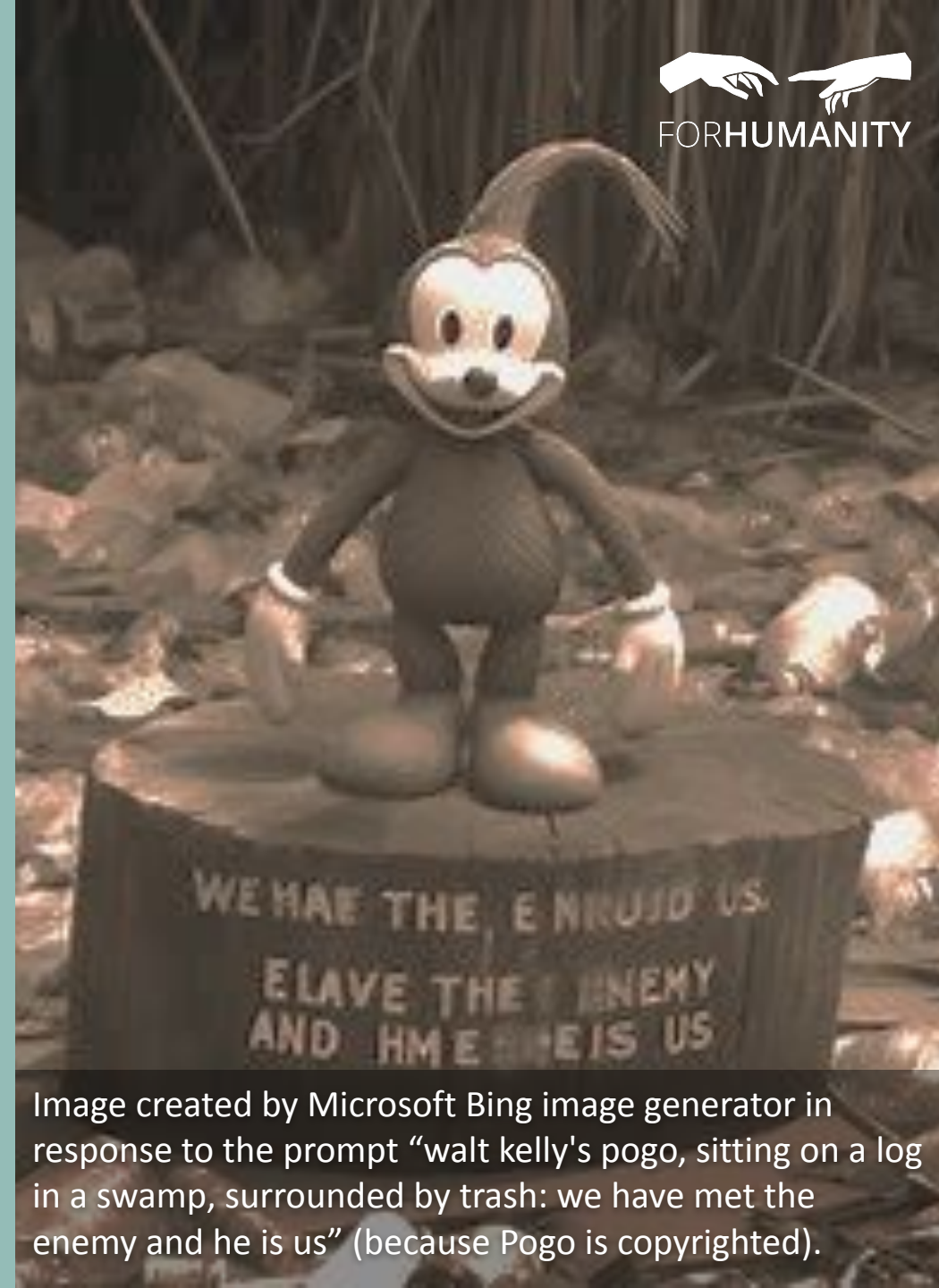


Image created by Microsoft Bing image generator in response to the prompt "walt kelly's pogo, sitting on a log in a swamp, surrounded by trash: we have met the enemy and he is us" (because Pogo is copyrighted).

How do we all all come to suffer from Automation Bias?

<SPOILER ALERT> It's a vicious spiral

1. It starts with us being all-too-human, by trying to solve problems
2. Which we so often do these days by using technology
3. Technology often solves the problem, so we tend to trust it
4. This becomes our default position and perspective
5. We get to the point where we don't even realize that we default to it



Technology is not neutral

**"Technology is neither good nor bad; nor is it neutral."
From Melvin Kranzberg's six laws of technology**



Humanizing AAA Systems can lead to overreliance

We are enamored with new technology:

- We assume that it is always good
- We assume that it is always right
- We fail to question and verify

We can come to trust it more than our fellow human beings



You Don't Have to Hate AI to Question AI

If you are impacting others with AI, you have a duty to question the process and the outputs

There are systems that you can trust because they are proven, like calculators and elevators

Develop a proper understanding and a healthy skepticism of AAA Systems before you impose (accidentally or not) its values upon others



3. Why Automation Bias Matters?

Because it harms people -
and that includes **you**

And that's more than reason enough

Photo by CottonBro Studio on Pexels



AI Makes Mistakes

We've proven it enough that it matters



AI is not magical
AI is not always right
AI can be wrong,
sometimes very wrong



Many applications of AI are immature

When we fail to question and verify AI results we set ourselves up for failure

There are many reasons why AI can be wrong

Understanding why will help you find the right balance



Understanding Why AI Makes Mistakes



People are not perfect, and people make AI

or even questionable data

ifies them

second . . .

“the technological deck has
been stacked in advance in
favor of certain social interests”

- Langdon Winner,
Do Artifacts have Politics?

Over-reliance on AI Can Be Used to Exploit

People can and do exploit AAA Systems



AAA Systems can be used to exploit people, to divide people, to harm society as a whole

The profits are privatized -
and narrowly concentrated

The losses are socialized -
and widely distributed

Society loses when AAA Systems only serve to increase the wealth divide

Automation Bias Impacts Society

It will impact in a lasting way, and not a good way



We must recognize the common harms that can come from Automation Bias:

- Breakdown of communities and interpersonal relationships
- Abdicating human decision-making to machines
- De-skilling humans
- Systemic societal risk (e.g., market crashes, widespread unemployment)

Pilots Refuse to Believe Their Own Eyes

Varig Flight 254 crash, September 1989



The pilots mis-entered autopilot direction as 270.0 degrees, west, instead of 027.0 degrees, north

Even though the pilots could see that they were flying into the setting sun to the west, they refused to believe that they were wrong - until they ran out of fuel and crashed in the jungle

Most survived, but many died due to injuries as the flight landed far into the jungle

<https://www.dailymotion.com/video/x7n5qtk>



20 Years Later, Pilots Can't Even Fly The Plane

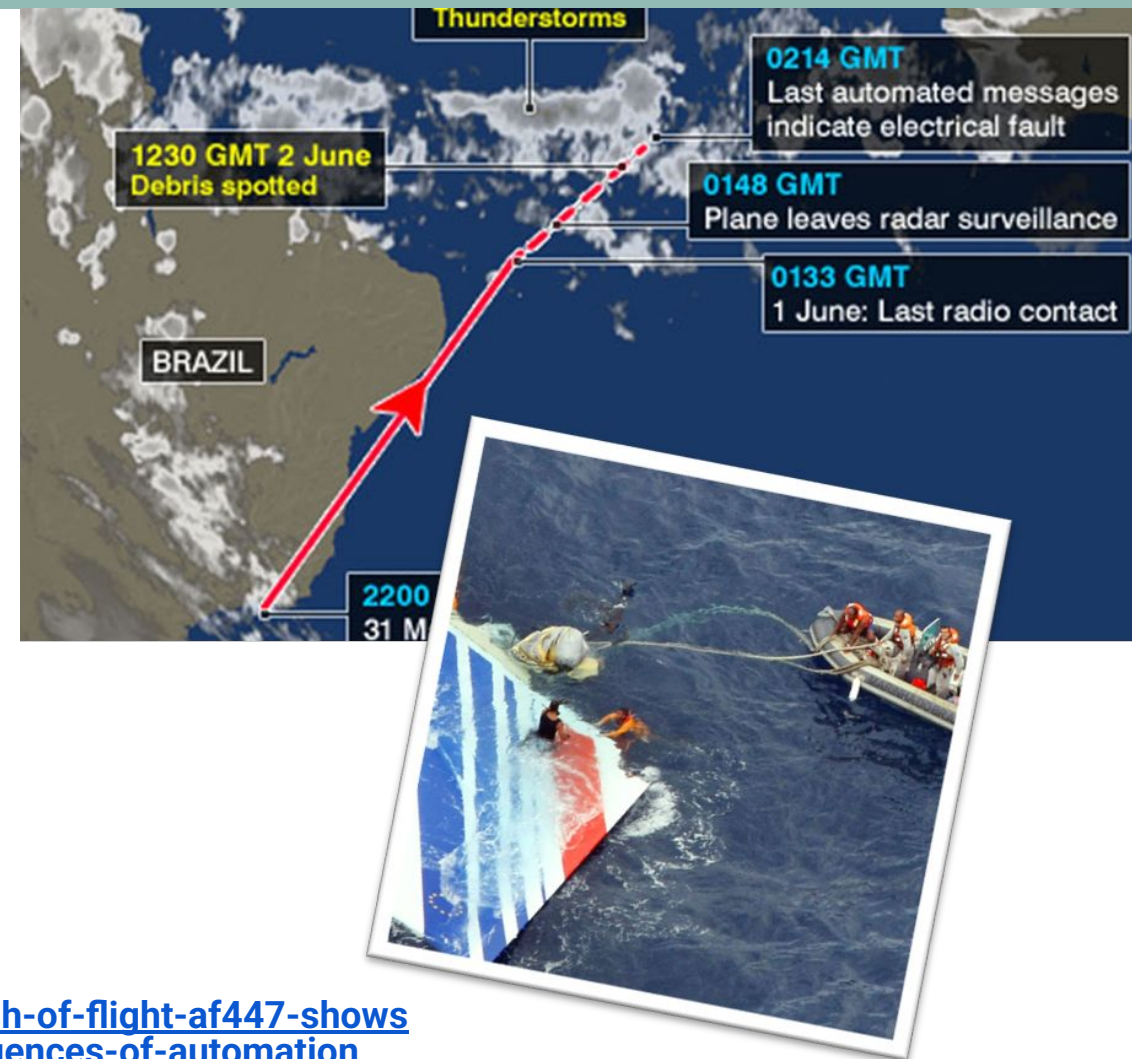
AirFrance Flight 477 crash, June, 2009



Automatic pilot suddenly disconnected during flight over the Atlantic, and the surprised pilot put the plane into a stall

The pilots did the *exact opposite* of what they needed to do to recover, sending 228 passengers and crew into a free fall into the ocean

Warning systems that had been designed to avoid distractions were shut down – the pilot's last words: "This can't be true"



<https://hbr.org/2017/09/the-tragic-crash-of-flight-af447-shows-the-unlikely-but-catastrophic-consequences-of-automation>

Now Might Be a Bad Time For Pilots to De-Skill

Deliberate GPS blocking - a new attack using AI Systems



“Commercial air crews are reporting something ‘unthinkable’ in the skies above the Middle East: novel “spoofing” attacks have caused navigation systems to fail in dozens of incidents since September.”

“... it’s not just GPS—fallback navigation systems are also corrupted, resulting in total failure.”

“... air traffic control had to step in and directly provide pilots ‘vectors’ (over an insecure communications channel) to guide them to their destination.”



<https://www.vice.com/en/article/m7bk3v/commercial-flights-are-experiencing-unthinkable-gps-attacks-and-nobody-knows-what-to-do>

Human oversight (trained experts) of AAA Systems remains one of our best forms of governance

AAA Systems Fall Short of Emulating Human Values

Based on programming, they ignore, repress, and actively suppress certain human values

If we are to live in a world filled with AAA Systems that replicate or replace human decision-making, then those decision-making processes should have the best aspects of our humanity embedded in the parameters.

Human interactions, based on values, are immensely complex and personal - Machines are not equipped for this type of engagement



AI Can Only Judge Us by Our History

Everyone has a right not to be judged solely by their worst day

Data can be:

- Incomplete
- Inaccurate
- Devoid of context

AAA Systems that use this data may:

- Act without context
- Never forget, unless we force them to do so
- Over-predict
- Create self-fulfilling prophecies
- Increasingly do this over generations (per Virginia Eubanks)

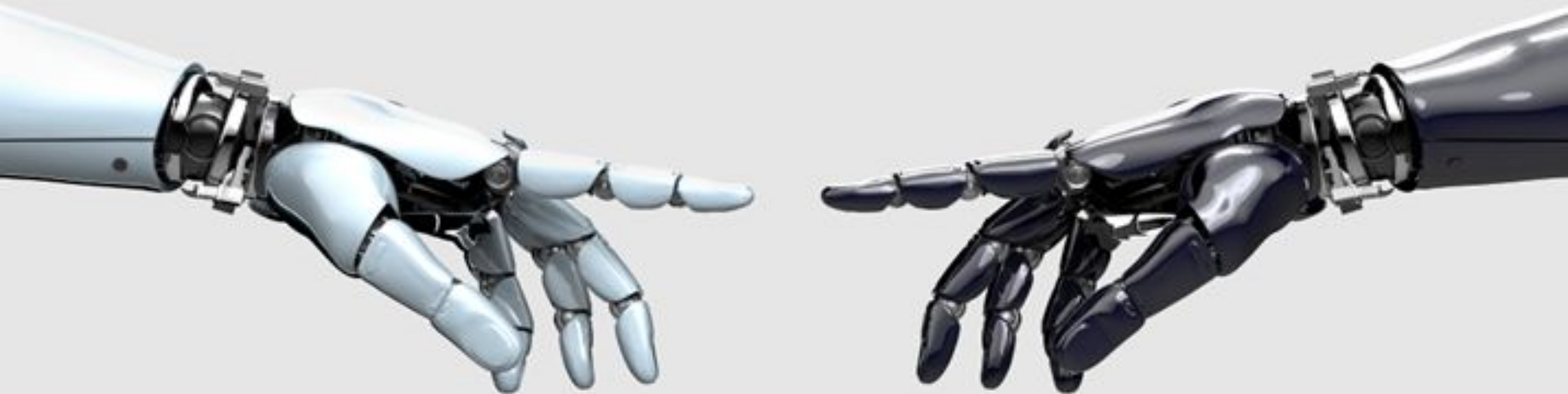


AAA Systems Amplify Existing Societal Issues

(e.g., bias, exclusion, discrimination)



And they create new and unintended consequences



“What’s Wrong With Stereotyping?”

AAA Systems categorize and classify us into little boxes

“. . . not all stereotypes are wrong in the sense of doing harm that is severe enough to warrant sanction from the law. . . stereotyping is wrong to the extent that it functions to deprive individuals of their freedom without good cause."

What’s Wrong with Stereotyping? Anita Bernstein, 55 ARIZ. L. REV. 655 (2013)

<https://arizonalawreview.org/whats-wrong-with-stereotyping/>



“How AI reduces the world to stereotypes”

Sometimes those little boxes come with pretty pictures -
and sometimes they *are* the pretty pictures



THE RISE OF AI

How AI reduces the world to stereotypes

Rest of World analyzed 3,000 AI images to see how image generators visualize different countries and cultures.

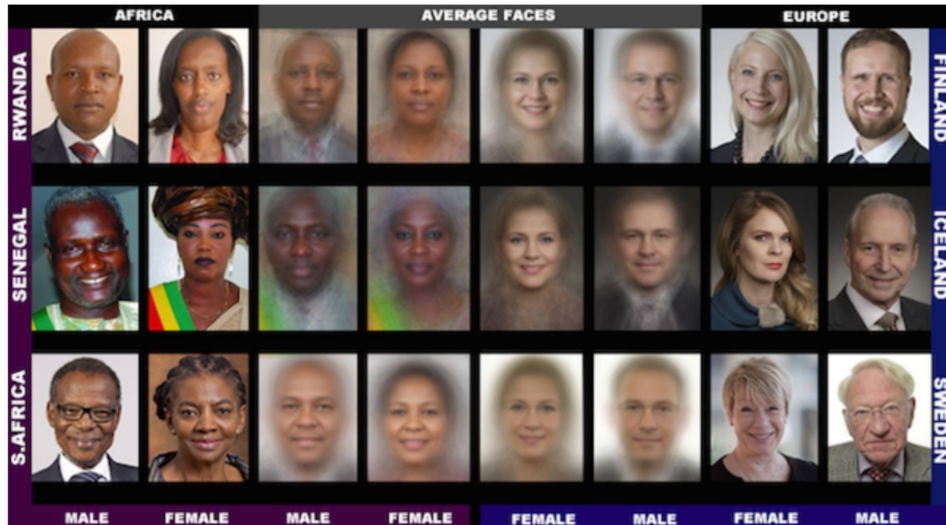
“A new Rest of World analysis shows that generative AI systems have tendencies toward bias, stereotypes, and reductionism when it comes to national identities, too.”

<https://restofworld.org/2023/ai-image-stereotypes/>

“A Mexican person” is usually a man in a sombrero.

Gender Shades is Still A Problem

“Leading tech companies' commercial AI systems significantly mis-gender women and darker skinned individuals”



“Companies do not disclose how well AI systems perform on different subgroups. Some admit to not checking.”

<http://gendershades.org/index.html>

Proceedings of Machine Learning Research 81:1–15, 2018

Conference on Fairness, Accountability, and Transparency

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini

MIT Media Lab 75 Amherst St. Cambridge, MA 02139

JOYAB@MIT.EDU

Timnit Gebru

Microsoft Research 641 Avenue of the Americas, New York, NY 10011

TIMNIT.GEBRU@MICROSOFT.COM

Editors: Sorelle A. Friedler and Christo Wilson

Abstract

Recent studies demonstrate that machine learning algorithms can discriminate based on classes like race and gender. In this work, we present an approach to evaluate bias present in automated facial analysis algorithms and datasets with respect to phenotypic groups such as the dermatological

who is hired, fired, granted a loan, or how long an individual spends in prison, decisions that have traditionally been performed by humans are rapidly made by algorithms (O’Neil, 2017; Citron and Pasquale, 2014). Even AI-based technologies that are not specifically trained to perform high-stakes tasks (such as determining how long someone spends in prison) are being used in pipelines

But Bias Can Begin Even at the Basics

Context can become a constraint



“All data is historical data: the product of a time, place, political economic, technical, and social climate. If you are not considering why your data exists, and other data sets don't, you are doing data science wrong.”

- Melissa Terras

For many AAA Systems:

Designers use only Western language and data (knowledge system)

Developers only source data and functionality based on that knowledge system

Deployers create systems that cannot read or interact with other knowledge systems

Users can only engage based upon the provided knowledge system

4. What Do I Need to Know/Do?

Now that you know what Automation Bias is and why it is a problem, the next step is to learn what to do about it.

How do we best recognize and rectify Automation Bias? And since it is clearly a big problem, how do we organize to be able to present solutions at scale?



AAA Systems Rarely Deal in Causality

It's all too easy to attribute false intentionality to AI



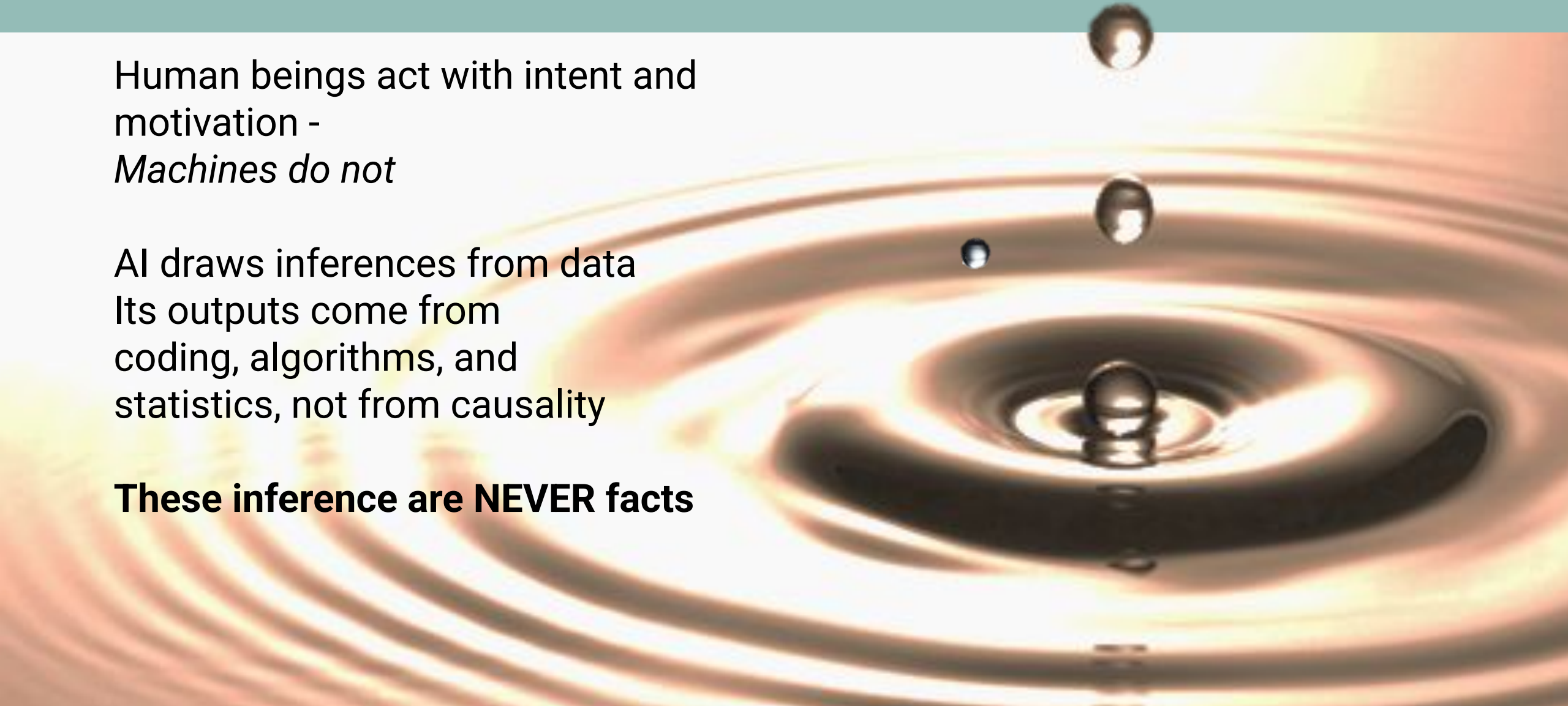
Human beings act with intent and motivation -

Machines do not

AI draws inferences from data

Its outputs come from coding, algorithms, and statistics, not from causality

These inference are NEVER facts



AI Thinks Your Previous Coin Flips Control Your Next One

AI gets stuck in probabilistic “thinking”

Humans can use probabilistic thinking - like anticipating a 30% chance of rain - but we don't turn every decision into a mathematical equation, like AI does.

- AI often overweights certain outcomes
- 7 HEADS in a row doesn't mean the next flip will be TAILS - this is Gambler's Bias
- An unbiased coin toss is independent of its history - the odds are always 50/50.
- Your odds, as a person with autonomy, are also independent of your history. You have the capacity to change. AI cannot.



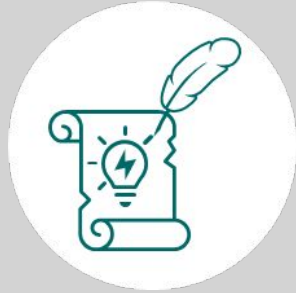
Conclusions From AAA Systems are NOT Facts

“Don’t trust, verify”



FACTS

Allow us to operate with certainty



Have a history of acceptance



Can be tested and proven



Acceptance is a cultural artifact

NOT FACTS

Cannot provide certainty



Hypothesis



Predictions



Inferences

How should We React to Output from AAA Systems?

AI outputs are inferences, not facts

Outputs from AAA Systems should **NOT** be treated as:

1. perfect
2. unquestionable
3. precise and correct
4. inevitable

We should verify outputs with fact checking or ground truth validation from separate sources



AAA Systems Should Not Be Blindly Trusted

Introduce a healthy balance between human judgment and AI suggestions



Educate on Limitations

Teach others how AI is trained on past data and may not accurately suggest actions in novel situations



Promote Critical Thinking

Encourage questioning, analyzing, and evaluating AI suggestions - rather than just accepting



Practice Decision-Making

Regularly make decisions without relying on AI - to develop confidence in human judgment and reduce over-reliance on AI



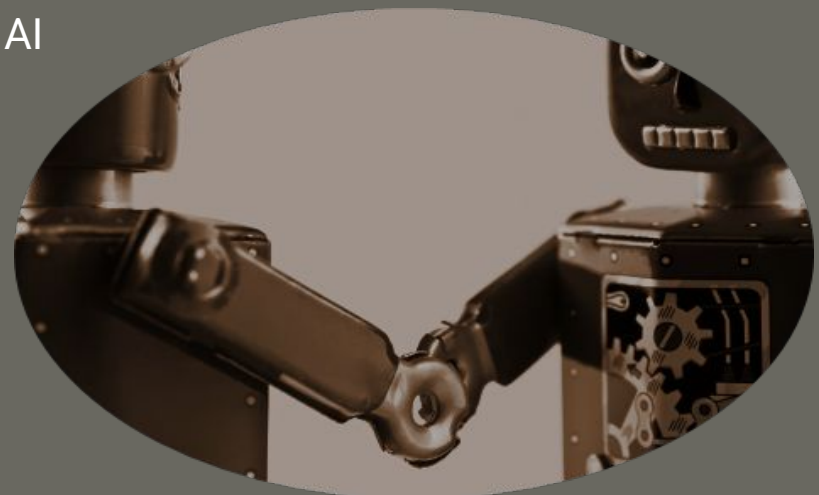
Encourage Skepticism

Leverage doubt as a powerful tool - promote the practice of cross-verifying information



Solicit Feedback

Use feedback on AI suggestions to improve the system, reduce errors, and increase reliability





Educate on Limitations

Question the training dataset of the system

Why? Human bias is repeated in many training datasets

- A large dataset of faces might have insufficient examples of protected categories or intersectionalities
Example: [Gender Shades](#) for Black females
- AI runs the risk of magnifying existing biases by embedding bias and expanding the impact
Example: [Amazon hiring tool](#) for gender bias in hiring

Be aware of potential bias in AI!





Educate on Limitations

Question the training dataset of the system

How?

- Know the scope, nature, context, and purpose of the tool
- Know where to look to find dataset descriptions
- Know what to look for how those datasets are constituted
- Know common bias in those datasets
- Consider how the labelling process was executed by the tool
- Seek academic analysis of the datasets

Consider all potential aspects of the dataset





Educate on Limitations

Question the training dataset of the system



However . . .

- Documentation lacks information on the diversity and bias of the data used so is difficult to identify what bias can manifest within the tool before using them
- Can be very difficult to find the documentation (it's often a very valuable trade secret)
- Many (e.g., ChatGPT and Dall-e) are trained using web data which itself can create bias against information, cultures, less common on the internet

Detecting and correcting AI bias is often easier said than done





Educate on Limitations

Question the training dataset of the system

Uncover the biases in your AAA systems

- Read AAA systems user guide, model cards, datasets for datasheets, data transparency document, technical documentation
- Examine incident databases - like <https://incidentdatabase.ai/>
- Research the method(s) used to train the model
- Perform your own or leverage existing human evaluations

Humans need to be aware themselves of the presence of bias when using AI tools



AI Should Not Be Oversold

AI sales and marketing also requires oversight

Far too often AAA Systems lack ethical oversight of the language, presentation, and assessment of performance and marketing

The FTC has repeatedly warned about trying to sell AAA Systems as “magical” - to eliminate overhype and unlimited scope of model applicability

We need to make our goal a more clearly defined scope, nature, context, and purpose for AAA Systems that includes disclosure of Residual Risks



Despite the Warnings, AI Hype is Endemic

But is it still good business?



BY TOM KRISHER

Updated 7:14 PM EDT, October 23, 2023

Share

DETROIT (AP) — Federal prosecutors have expanded investigations into Tesla beyond the electric vehicle maker's [partially automated driving systems](#), and they have issued subpoenas for information instead of simply requesting it, the company disclosed Monday.

In a quarterly report filed with the Securities and Exchange Commission, Tesla said the Department of Justice is looking into "personal benefits, related parties, vehicle range and personnel decisions," without giving details.

The additional investigation topics and the subpoenas suggest that prosecutors have broadened their inquiry, and they have found the need to force Tesla to disclose information, legal experts say. The filing indicates prosecutors may be investigating Tesla CEO Elon Musk, and whether the company has been candid in describing the features of its vehicles, they say.

<https://apnews.com/article/tesla-investigations-justice-department-musk-self-driving-29a68864f75c9fabbd04f7a87d169444>



Tesla crashes at high speed into Greater Columbus Convention Center

Video of a violent car crash at the Greater Columbus Convention Center May 4 shows the Tesla ran a red and smashed into the side of the building. *The Columbus Dispatch*

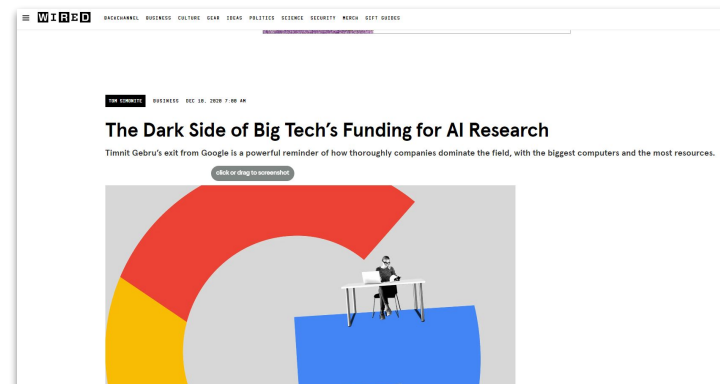
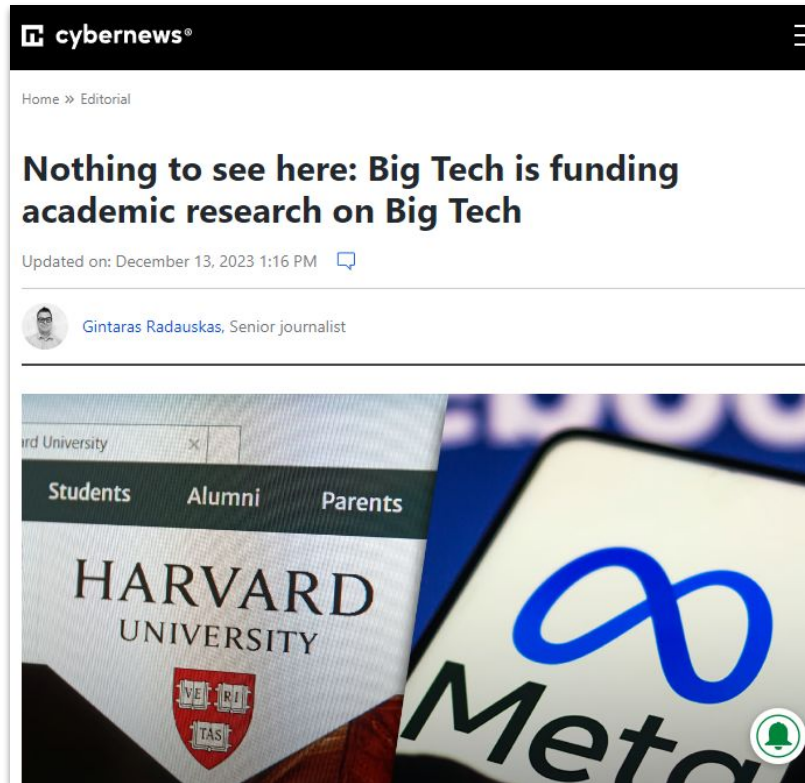
<https://www.dispatch.com/story/news/2022/05/11/tesla-crash-into-columbus-convention-center-watch-video-release-d/9731137002/>

Beware of Headline Hype - Check the Facts

Much of today's academic research is funded by Tech companies



Overcoming Automation Bias requires us to independently evaluate claims and avoid assuming legitimacy based on legacy perceptions



Trust is Earned

It cannot be bought - and there are no substitutes for time and experience

How and why does technology gain cultural acceptance?

It takes time

It takes experience (whether direct or indirect/societal) with the safety features - that they work, that they are there

So, when you approach new technology, it is fair (and even wise) to approach it with a degree of skepticism



It's a Function of Risk

The more risky the AAA System is to humans, the greater our duty to overcome Automation Bias

We cannot treat all potential risks from AAA Systems the same. That would be exhausting, even if we had the necessary resources.

Instead, we need to take a risk-based approach. We need to evaluate the likelihood of harm and the impact of the harm - and weigh those together to prioritize our actions.



Risk Rating Matrix - Impact

Severity Model - specification/hierarchical comparison From highest to lowest



1. Catastrophic	Life/Death Decisions Physical Or Mental Harm Loss Of Rights/Freedoms
2. Major	Restrictions Of Rights/Freedoms Restriction To The Access Of Goods And Services Discriminatory Outcomes
3. Moderate	Unfair Outcomes Identify Theft/Loss Of Identity Disclosure Of Personal Information
4. Minor	Damage To Reputation Monetary Loss Harassment And Increased Undesirable Exposure
5. Insignificant	Annoyance And Hassle Minor Repairs Petty Disturbance

Critical aspects of impact across risk inputs:

- Magnitude
- Relativity
- Duration
- Geographic scope

Risk Rating Matrix - Likelihood

Probability Model - specification/hierarchical comparison From highest to lowest



1. Near Certain (greater than 95% chance)
2. Very likely (70%-95%)
3. Likely (50%-70%)
4. Possible (30%-50%)
5. Unlikely (15%-30%)
6. Very unlikely (5%-15%)
7. Remote (1% to 5% chance)
8. Rare or Unusual (less than 1% chance)

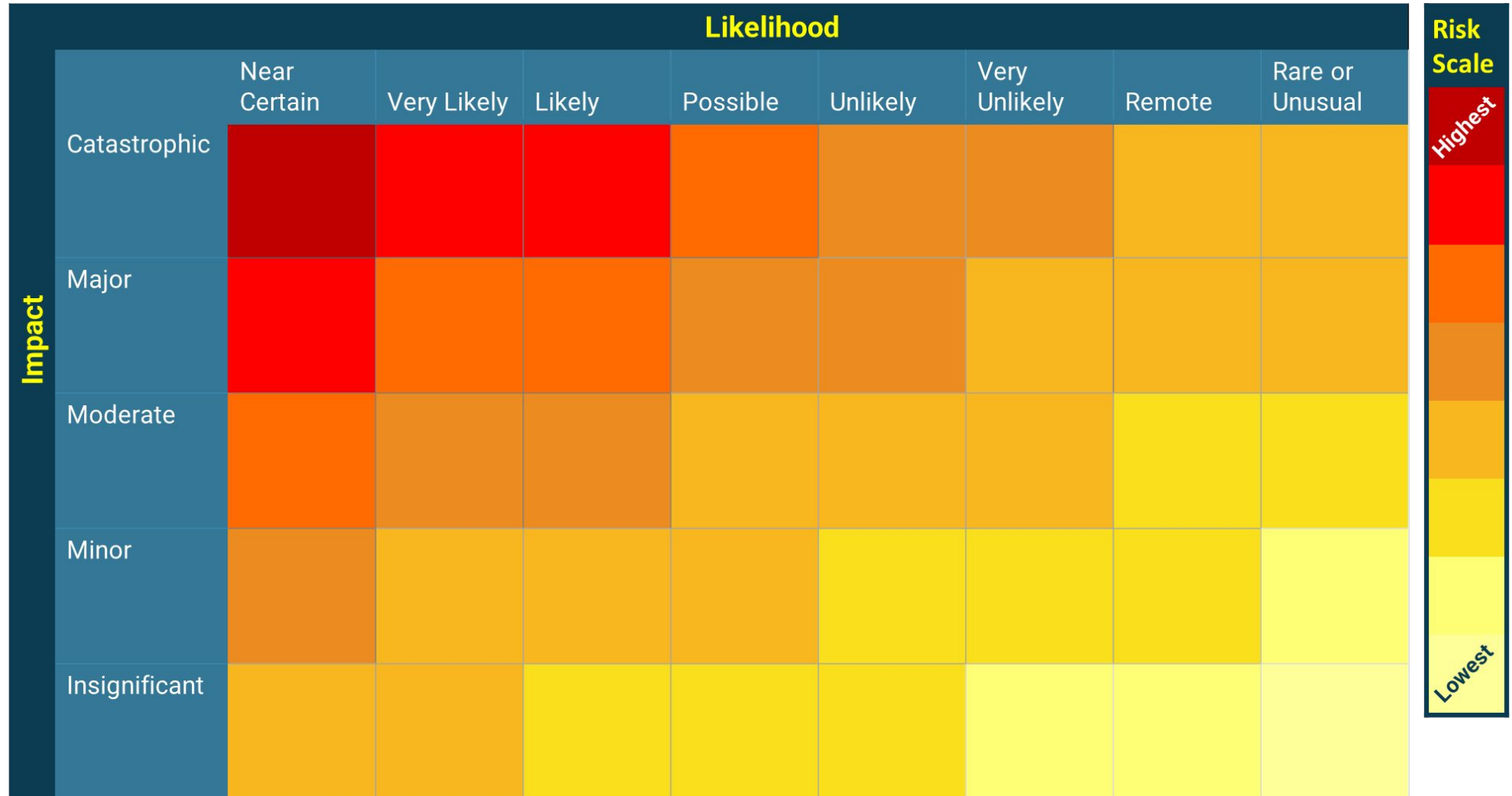
CRITICAL QUESTION: Have we changed our safety protocols according to likelihood or has likelihood changed because of previous risk assessment and mitigation?

EXAMPLE: the airline industry,
one death per 750,000 boardings during 1978-1987
one death per 7.9 million boardings in 2017

Designating the Deeply Scary Danger Zones



While this chart displays the most important areas of concern, it does not display how those risks should be properly weighted - along with the tail risks



Likelihood Adjusted for Volume

There are far more potential Rare/Unusual events than Near Certain ones



1. Near Certain (greater than 95% chance)
2. Very likely (70%-95%)
3. Likely (50%-70%)
4. Possible (30%-50%)
5. Unlikely (15%-30%)
6. Very unlikely (5%-15%)
7. Remote (1% to 5% chance)
8. Rare or Unusual (less than 1% chance)



Dangerous “Black Swan” events (negative occurrences believed impossible or not to exist, subsequently shown to exist) lurk within the “Rare or Unusual” area

Per Nassim Nicholas Taleb

Roy Muz on Unsplash

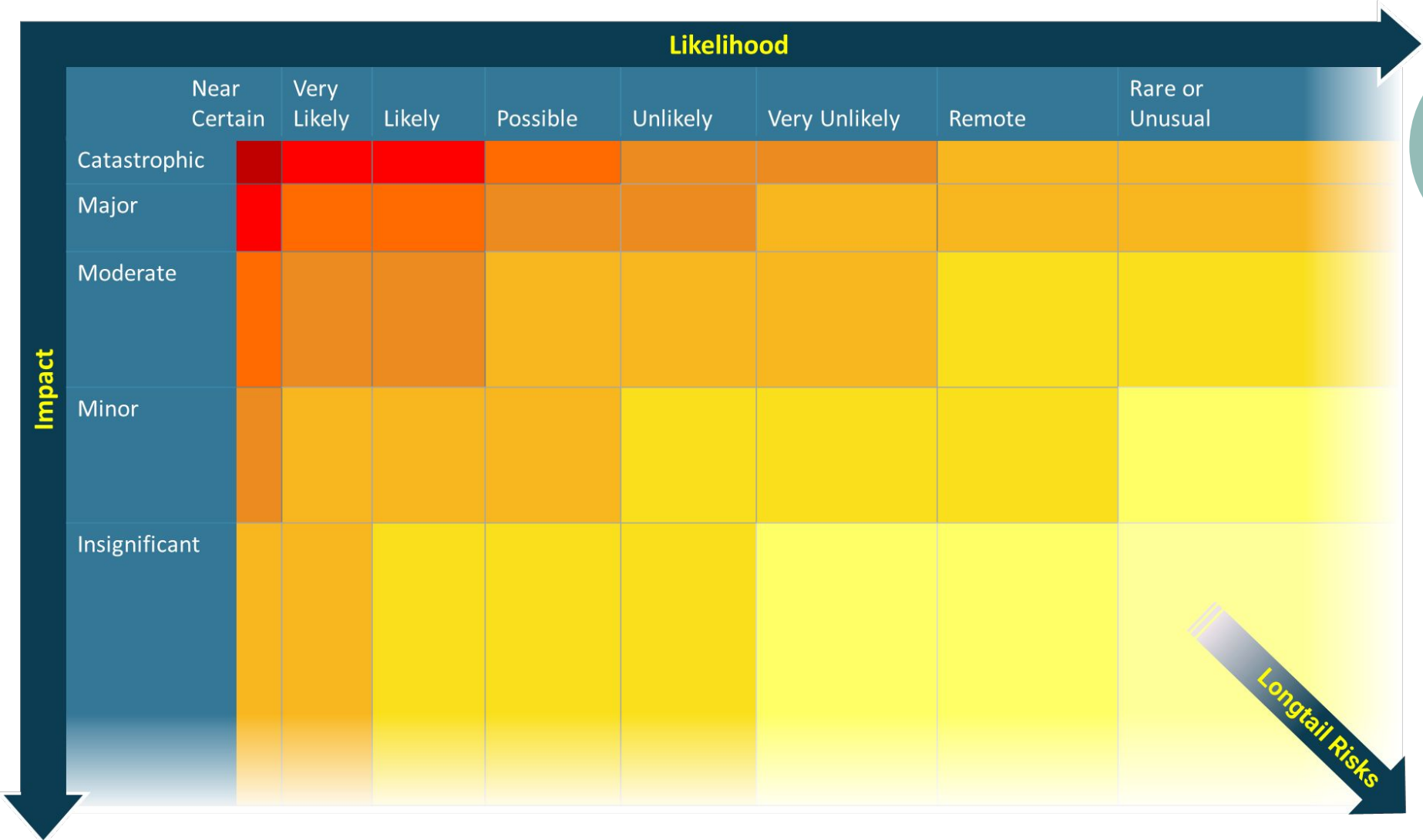
Humans regularly underweight statistically improbable events
Something that you think is rare or unusual might not be nearly as rare or unusual as you think
Your Black Swans are probably just something that was underweighted in the first place

Risk Rating Matrix - Likelihood

Scaled to show the volume of events and the longtails at the margins



The next adjustment to the chart shows the properly weighted risks and the potential for tail risks and Black Swan events



Black Swans

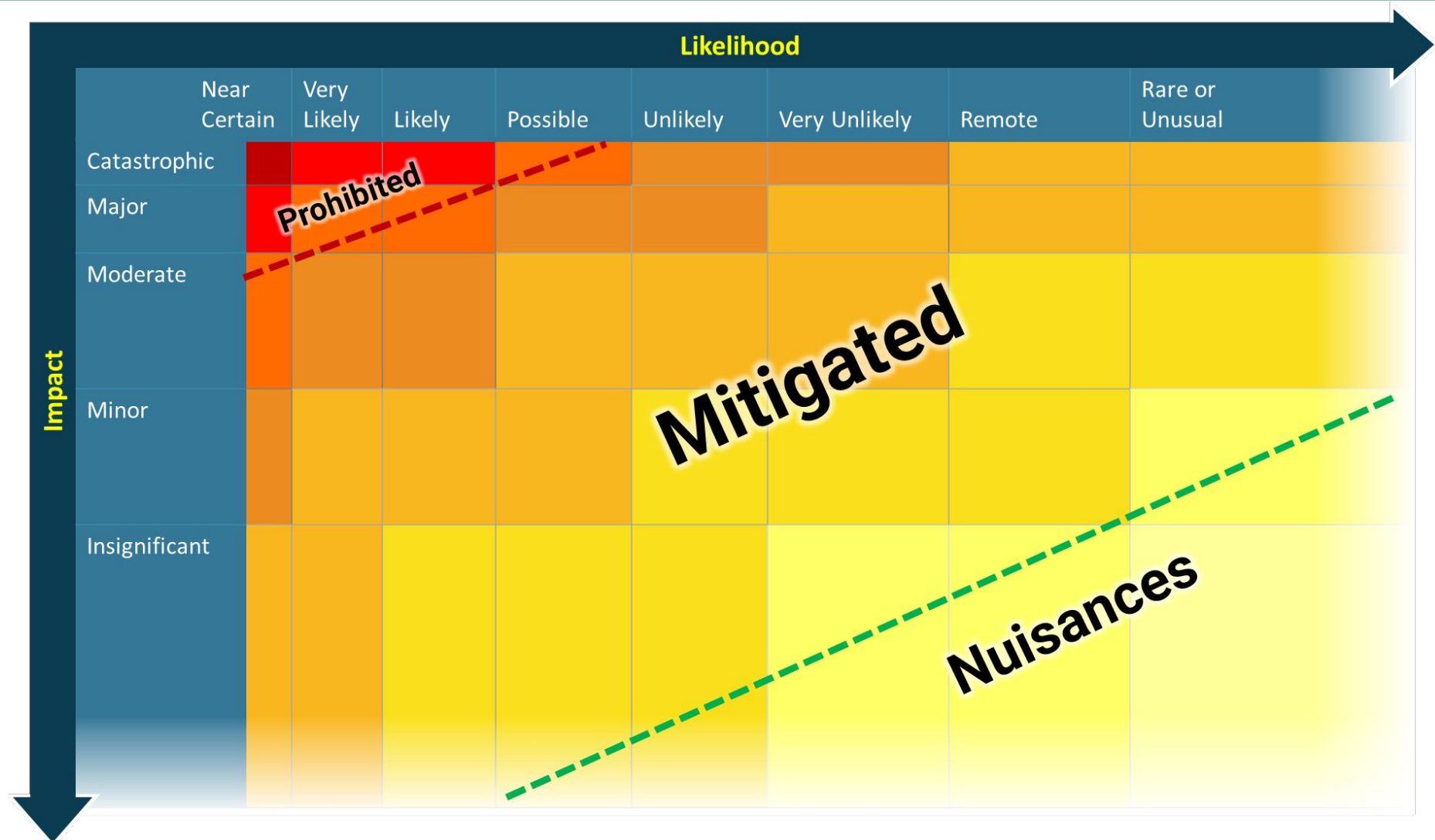
Risk Rating Matrix - Categorization

Not all risks should be treated the same way, but they can be grouped



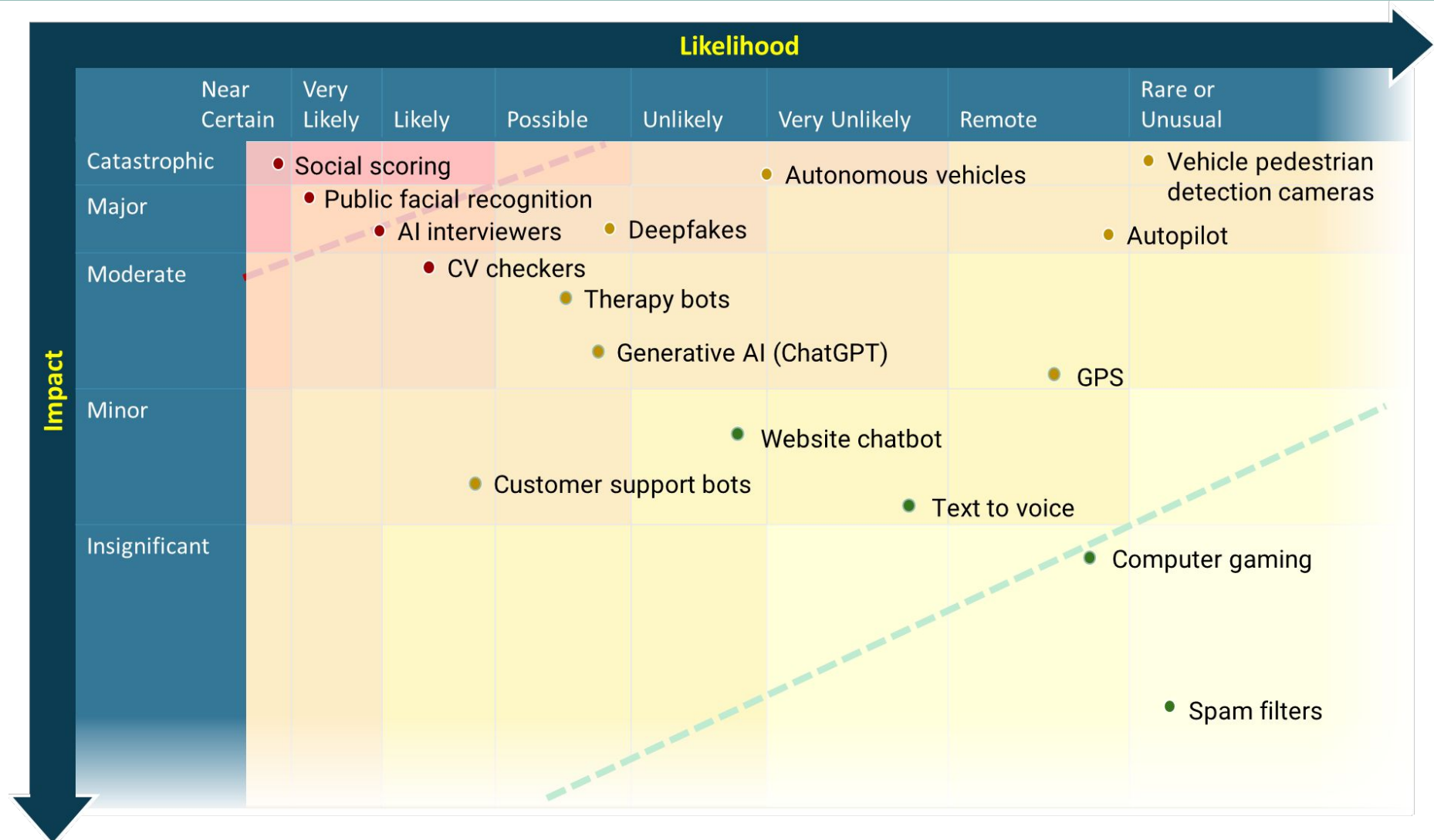
Risks can be generally grouped into 3 categories according to how we need to deal with them:

1. Prohibited
2. Mitigated
3. Nuisances



FORHUMANITY

The final step is to score the individual risks on the chart



Methods for Mitigating Bias

Consider bias in all of the forms it may manifest

Consider Bias in at least three forms:

- 1) **Statistical Bias** - Are the data inputs representative?
- 2) **Cognitive Bias** - Does the manner in which data is collected result in bias (e.g., if a dropdown box only offers 4 ethnicities, those choices will generate bias)?
- 3) **Non-Response Bias** - Does the technology being deployed prohibit some users from access?

These forms of bias may manifest themselves in:

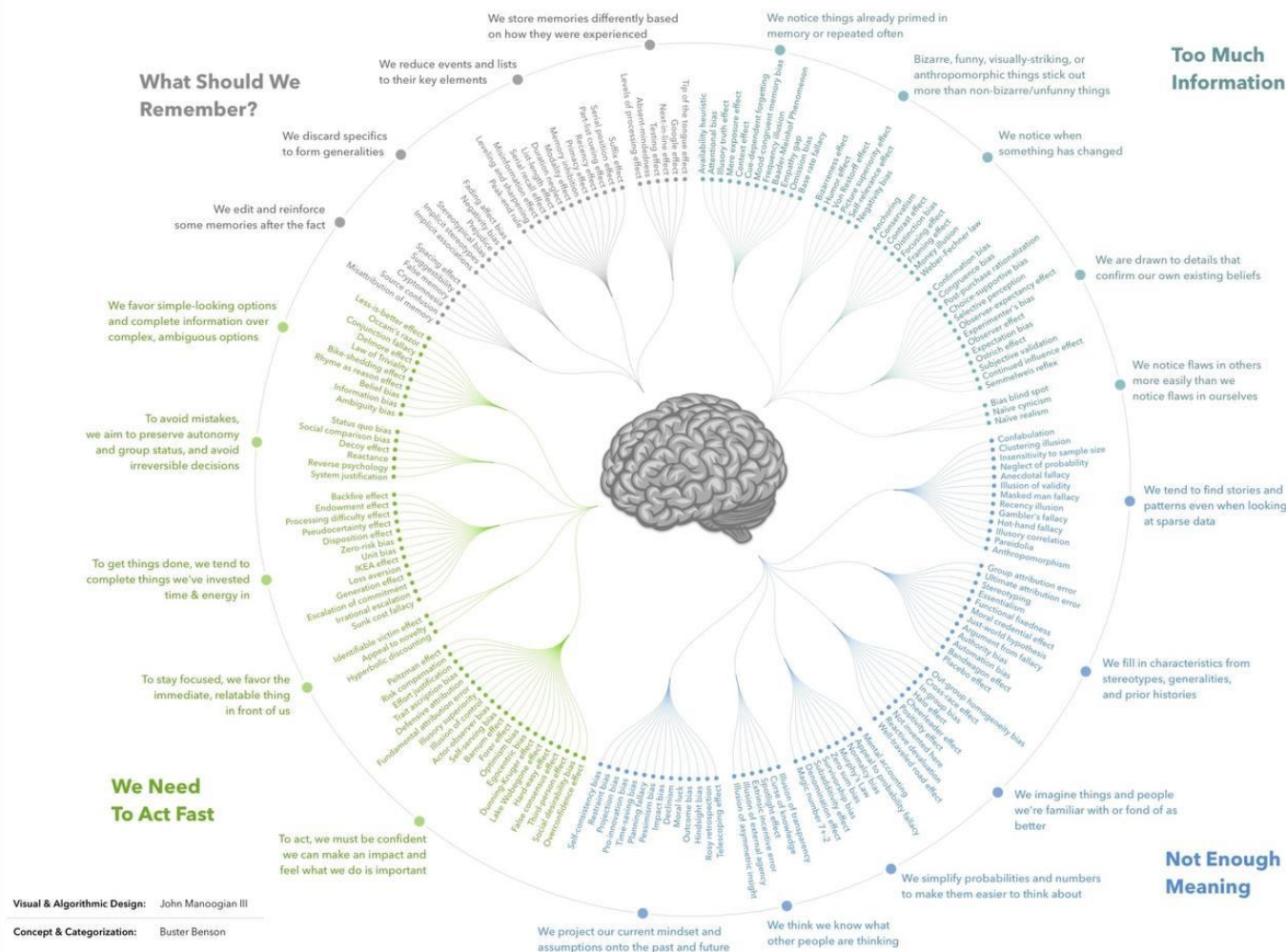
- 1) **Data** - Understand the nature and characteristics of the data
- 2) **Architectural Inputs** - Understand how feature selection and weightings might result in bias
- 3) **Outcomes** - Test outcomes against selection metrics for disparate impact





FOR HUMANITY

But it is necessary



Concept & Categorization: Buster Benson



There Are Over 130 Listed Biases

These are just a few (from https://en.wikipedia.org/wiki/List_of_cognitive_biases)



Agent detection bias, the inclination to presume the purposeful intervention of a sentient or intelligent agent.



Anchoring bias: the tendency to rely too heavily—to "anchor"—on one trait or piece of information when making decisions.



Anthropomorphism: the tendency to characterize animals, objects, and abstract concepts as possessing human-like traits, emotions, and intentions.



Apophenia: the tendency to perceive meaningful connections between unrelated things.



Bias blind spot: the tendency to see oneself as less biased than other people, or to be able to identify more cognitive biases in others than in oneself.



Clustering illusion: the tendency to overestimate the importance of small runs, streaks, or clusters in large samples of random data.



Conjunction fallacy: the tendency to assume that specific conditions are more probable than a more general version.



Confirmation bias: the tendency to search for, interpret, focus on and remember information in a way that confirms one's preconceptions.



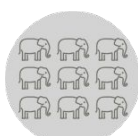
Conservatism bias: the tendency to insufficiently revise one's belief when presented with new evidence.



Default effect: the tendency to favor the default option when given a choice between several options.



Dunning-Kruger effect: the tendency for unskilled individuals to overestimate their own ability.



Frequency illusion: that once something has been noticed then every instance of that thing is noticed, leading to the belief it has a high frequency of occurrence.



Gambler's fallacy: the tendency to think that future probabilities are altered by past events, when in reality they are unchanged.



G.I. Joe fallacy: the tendency to think that knowing about cognitive bias is enough to overcome it.



Illusory correlation: a tendency to inaccurately perceive a relationship between two unrelated events.



Illusory truth effect: the tendency to believe that a statement is true if it is easier to process, or if it has been stated multiple times, regardless of its actual veracity.



Impostor syndrome: a psychological occurrence in which people doubt their skills, talents, or accomplishments.



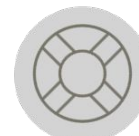
Law of the instrument: an over-reliance on a familiar tool or methods, ignoring or under-valuing alternative approaches.



Normalcy bias: the refusal to plan for, or react to, a disaster which has never happened before.



Overconfidence effect: a tendency to have excessive confidence in one's own answers to questions.



Survivorship Bias: if you ask only the lottery winners about the lottery, they will tell you that it's great and nobody loses.

5. How to Recognize Automation Bias

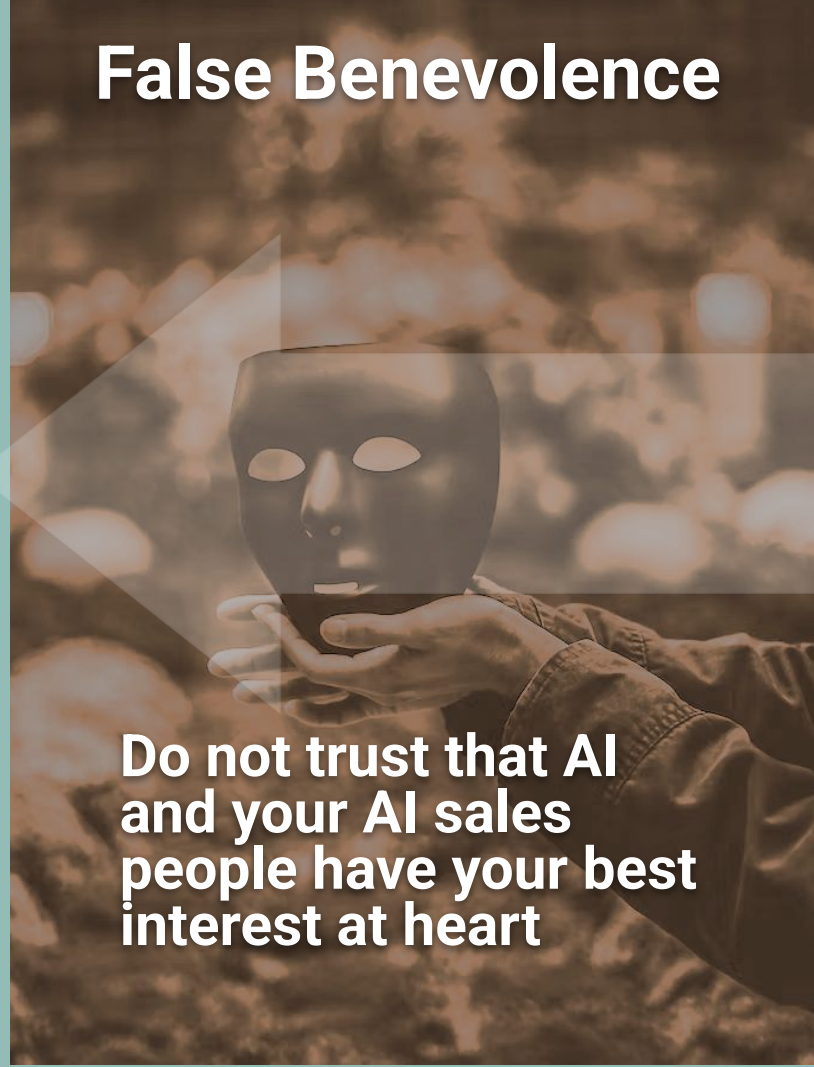
Recognizing the cyclical nature of most Automation Bias might be at or near an all-time high, so we need to recognize that there might be a societal-wide shift that reduces Automation Bias and the risks that arise from it



Environments where Automation Bias presents



False Benevolence

A hand is shown holding a black, featureless mask. The background is a warm, bokeh-style image of light spots. A large, semi-transparent white arrow points from the right towards the mask.

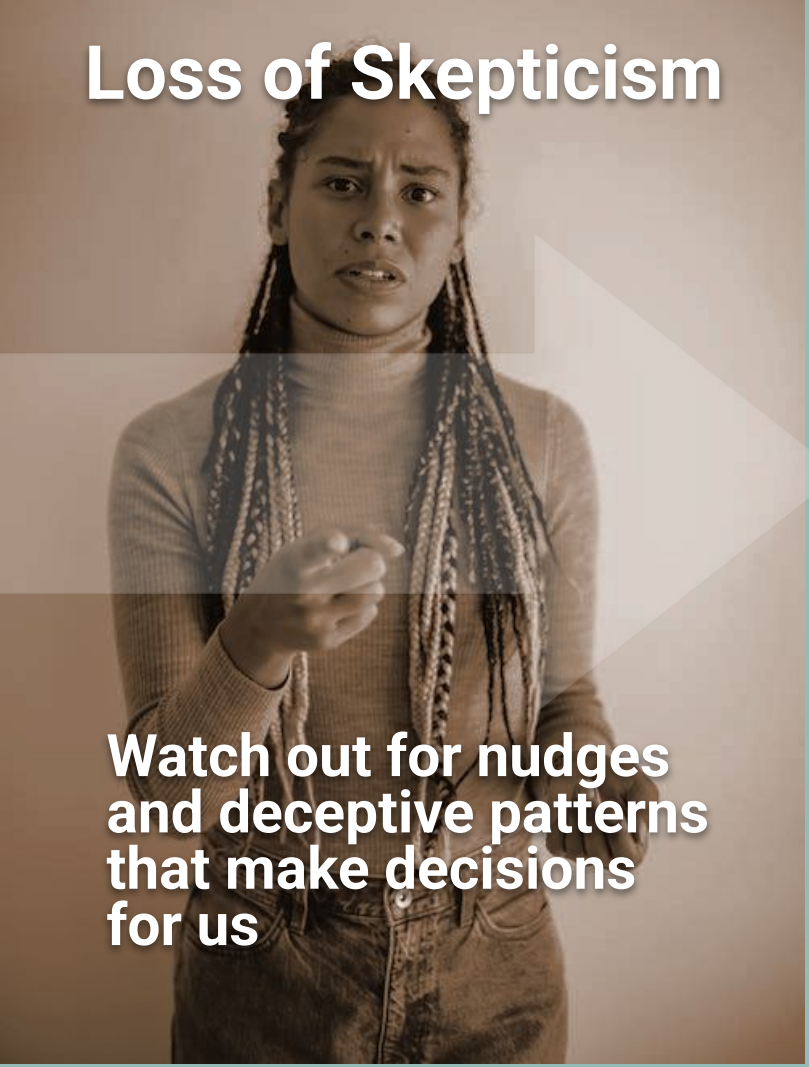
**Do not trust that AI
and your AI sales
people have your best
interest at heart**

Over-trust

A close-up shot of many hands reaching out from a crowd, creating a sense of collective action or support. A large, semi-transparent white arrow points from the left towards the hands.

**Do not place excessive
trust in technology,
require technology to
earn your trust**

Loss of Skepticism

A woman with long braids, wearing a beige turtleneck and a patterned scarf, looks directly at the camera with a concerned expression. A large, semi-transparent white arrow points from the right towards her.

**Watch out for nudges
and deceptive patterns
that make decisions
for us**

Situational Awareness for Automation Bias

If you don't remove your rose-colored glasses, you'll never see that some flags are red



New technology

Old technology
used without
appropriate due
diligence

Assumption that
higher authority
has vetted or
approved the
technology

Trusted brand,
endorser, or
influencer

Hyper-
competition that
values speed
over safety

Trust premised
on paid or
premium
product/service
over free

Open source or
peer review

National pride in
product or
creator

UX/UI nudges
and deceptive
patterns

Situational Awareness for Automation Bias

If you don't remove your rose-colored glasses, you'll never see that some flags are red



New tech doesn't automatically = better results

Provide adequate user training and guidance

Few rules currently exist to govern or vet these technologies.

Spokespeople and brands have their own agendas

Prioritize safety as a user to help combat this bias

If a service is free, assume you are the product.

NOT SURE WHAT THIS ONE MEANS??

Big Tech is global. Don't assume 'made in your country' = more trustworthy

Be aware of common deceptive patterns.

Symptoms and Contributions to Automation Bias

“If any these symptoms should persist, consult your AI ethicist immediately”



Ignoring your gut
instinct

Fail to check
sources and
citations

Unthinking use
of AAA Systems

Assuming
correct or
accurate
outcomes

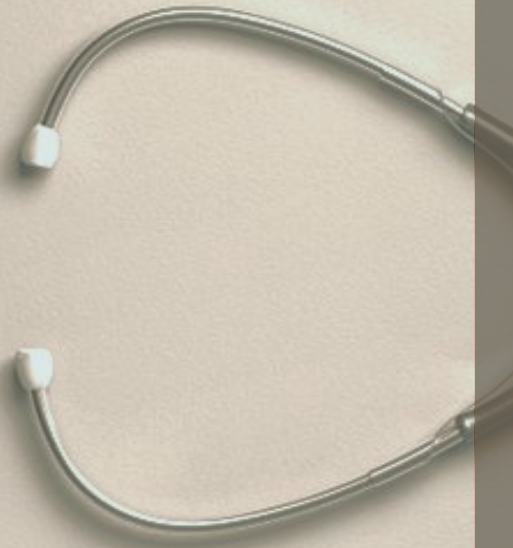
Failure to
appreciate
frequency of
inaccurate
outcomes

Unawareness of
negative or
failed outcomes

Belief in the
machine over
belief in the
human without
evidence

Steadfast
adherence to
machine
decisions
despite evidence

Different
standards
applied to
machines versus
humans



Symptoms and Contributions to Automation Bias

“If any these symptoms should persist, consult your AI ethicist immediately”



Trust your gut
instinct

Apply critical
thinking -
Check sources
and citations

Use internal
questions and
comment tools
to highlight
concerns

Don't assume the
system provides
correct or
accurate
outcomes

Understand that
AAA Systems
fail regularly

Understand the
need for
appropriate
oversight and
intervention

Don't blindly
believe in the
machine over
human without
evidence

Be willing to
change your
mind when
presented with
evidence

Use appropriate
standards to
evaluate
human-machine
interaction

Recognize That You Have Choices

You don't have to "Google" everything



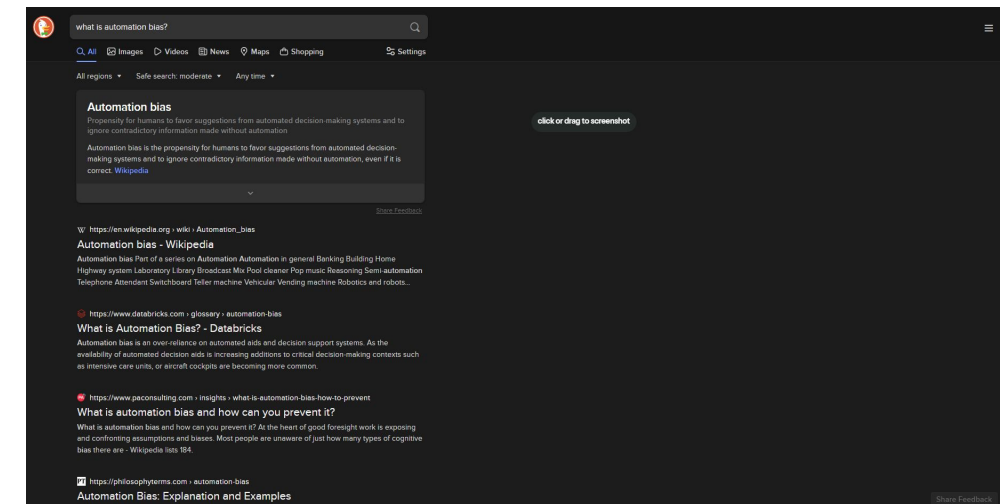
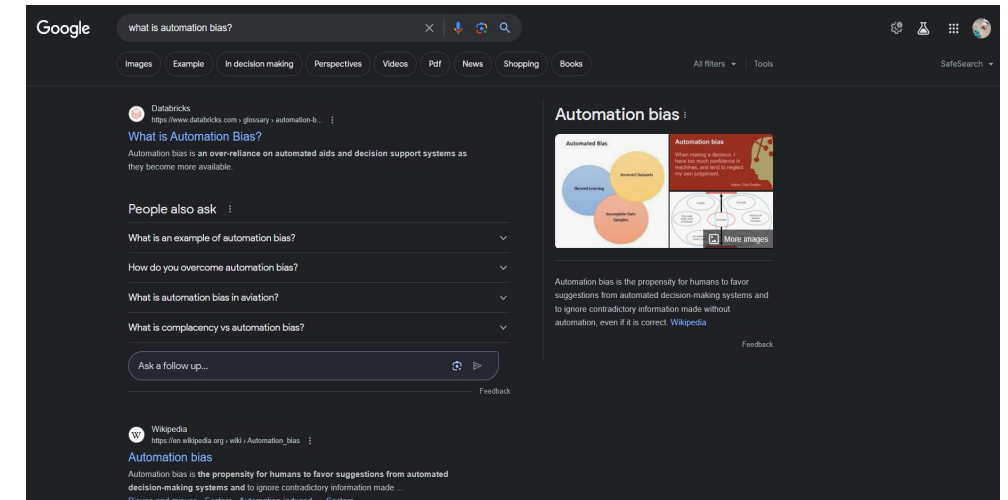
Alternatives exist

Take the time to find them

You could start by "Googling" them

But the better approach is to instead:

- Use another tool
- Ask another person/expert about your question; or
- Access other verified sources such as scholarly databases and peer reviewed journals



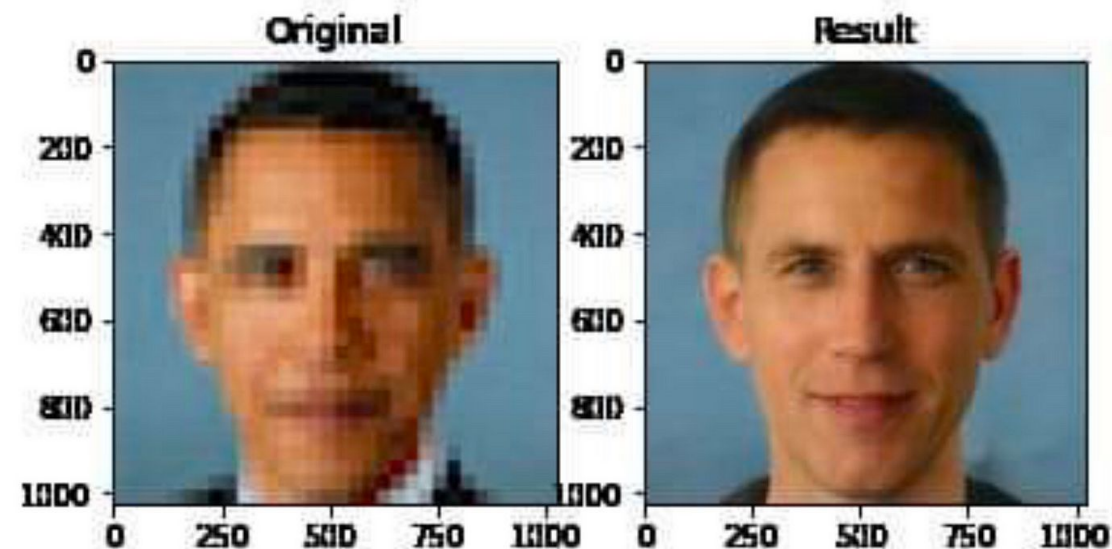
Skills and strategies:

Expect common bias problems - don't be surprised



Be prepared for bias in all of these areas - and more:

1. Race
2. Gender
3. Age
4. Socioeconomic
5. Culture (Western thought and philosophies)
6. Disability
7. Marital status
8. Sexual orientation



<https://towardsdatascience.com/how-are-algorithms-biased-8449406aaa83>

Skills and strategies:

Understand the bias behind the systems



INSIDER

Newsletters Log in [Subscribe](#)

HOME > TECH

An Asian MIT student asked AI to turn an image of her into a professional headshot. It made her white, with lighter skin and blue eyes.

Sawdah Bhaimiya Aug 1, 2023, 7:00 AM EDT



Why did an AI systems make this Asian woman white? Remember our discussion about data bias:

- How might the training data over index on white people?
- What data might have been labelled 'professional'? how might her prompt have been interpreted?
- We don't know for sure because the system is opaque

Expect Bias - Specific high-risk applications where we want to be vigilant



- Crime
- Bank Lending
- Employment

<https://www.forbes.com/sites/forbeshumanresourcescouncil/2021/10/14/understanding-bias-in-ai-enabled-hiring/?sh=4b14292c7b96>

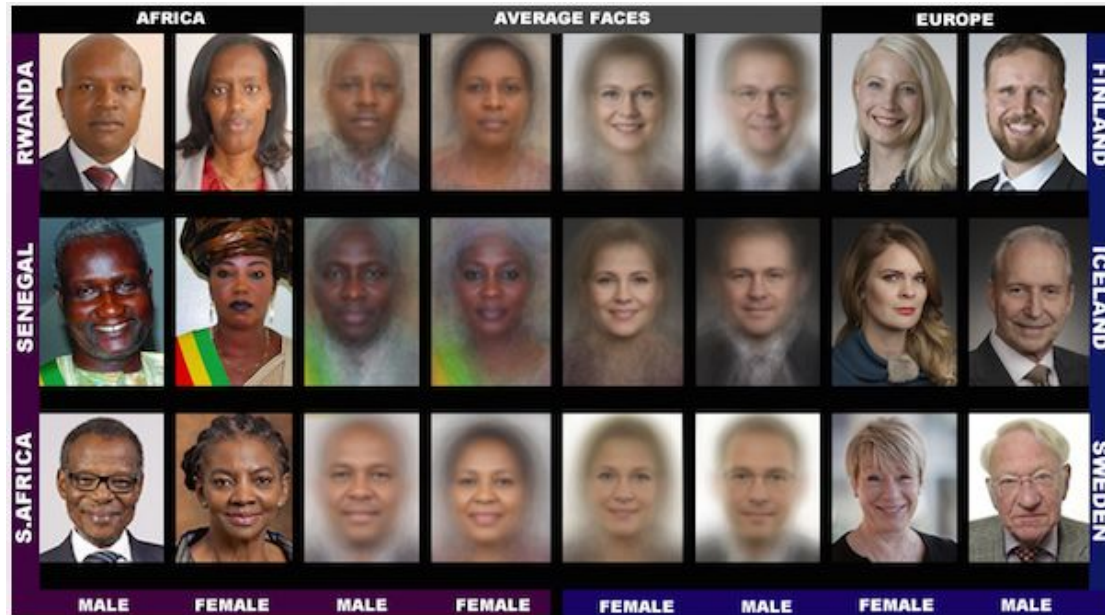
- Benefits- MiDAS/Rotterdam/
- Accounting - UK Post Office - Horizon

Skills and strategies:

Be prepared for feedback loops of bias re-entered into the system



Biased training data: the Gender Shades example



Biased tagging and processing: what's the default

(c) A **yellow** Vespa parked in a lot with other cars.



	Human Label	Visual Label
Yellow	✓	✓

(d) A store display that has a lot of bananas on sale.



	Human Label	Visual Label
Yellow	✗	✓

Credit: Data Feminism, introduction, p. 16

Pre-awareness of the Usefulness of an AAA System

Understand a system's strengths and weaknesses

- Study terms and conditions
- Understand the data behind the training/testing/validations dataset
- Understand the exceptions, anomalies, and dysfunctions
- Study Incident Databases (such as AI Incident Database)
- Tracking media for affirmations and negations of good usage



6. How do I Overcome Automation Bias

It's not enough just to understand and recognize Automation Bias

You need to act, and the best place to act, to begin the journey towards overcoming Automation Bias is by starting with yourself.



Our Capacity to Change Helps Make Us Human

AAA Systems fail to celebrate our ability to change, thus trapping people

Turning points can be the most important points in our lives:

- Redemption
- Forgiveness
- Ability/right to change

The very nature of these tools is to measure, analyze, and through correlation, explain your future using only your past. There is no turning point, no allowance for change; there is only the next step in the same direction.



Celebrate Human Values

**We cannot allow the systems meant to help us
make decisions make those decisions for us**

To be human and celebrate the best we have
in our collective spirit, we cannot allow AAA
Systems to ignore the best of our human
values such as:

- Collaboration
- Compromise
- Being self-less
- Kindness and Caring

Choosing to reject this fate begins with
combatting Automation Bias.

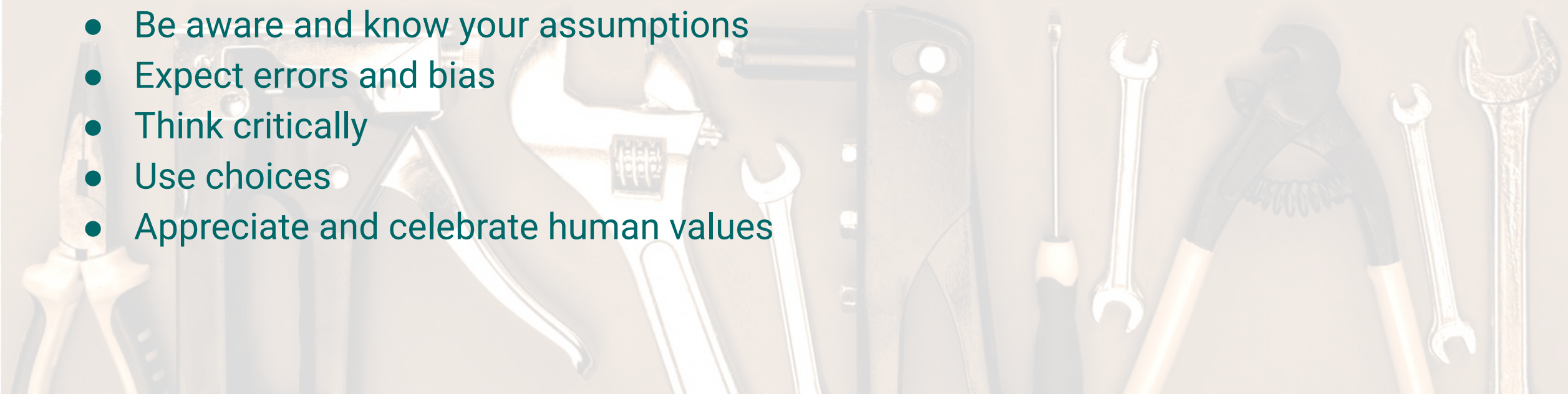


Knowledge, Tools, and Awareness

You need more than a hammer when not every problem is a nail



- Ask
 - “why”
 - about the cost/benefit analysis
 - “what are the consequences”?
- Education and training
- Be aware and know your assumptions
- Expect errors and bias
- Think critically
- Use choices
- Appreciate and celebrate human values

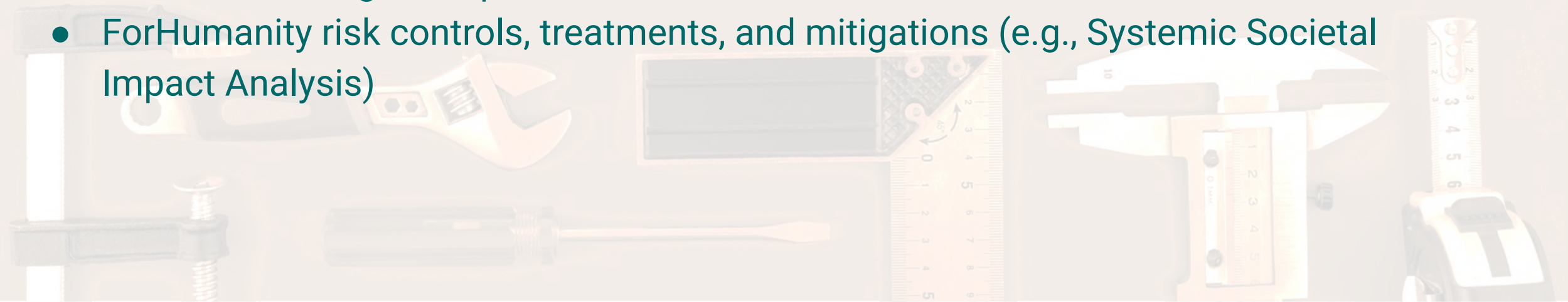


ForHumanity Tools

We have a plethora of powerful tools for you to use



- Algorithmic Risk Assessment
- Ethical Risk Assessment
- Necessity Assessment
- Proportionality Study
- Bias Risk Assessment
- Data Protection Impact Assessment
- Fundamental Rights Impact Assessment
- ForHumanity risk controls, treatments, and mitigations (e.g., Systemic Societal Impact Analysis)



Ask Yourself Why

Start with why

- Why am I considering using an AI?
- What is the purpose of using an AI?
- Are there are other solutions that are superior?
- Are there other solutions that uphold human dignity or skills?



Test Your Cost/Benefit Analysis

Do the results support automation and my shared moral framework (or is it all just about profit)?

- Who accrues the benefits?
 - Only shareholders?
 - Only Top Management and Oversight Bodies
 - What do the AI subjects get?
- Who bears the burdens or costs?
 - Loss of human jobs
 - Loss of human skill
- What tips the scales of your decisions?
 - Savings
 - Efficiency
 - Robustness/resilience
- Proportionality

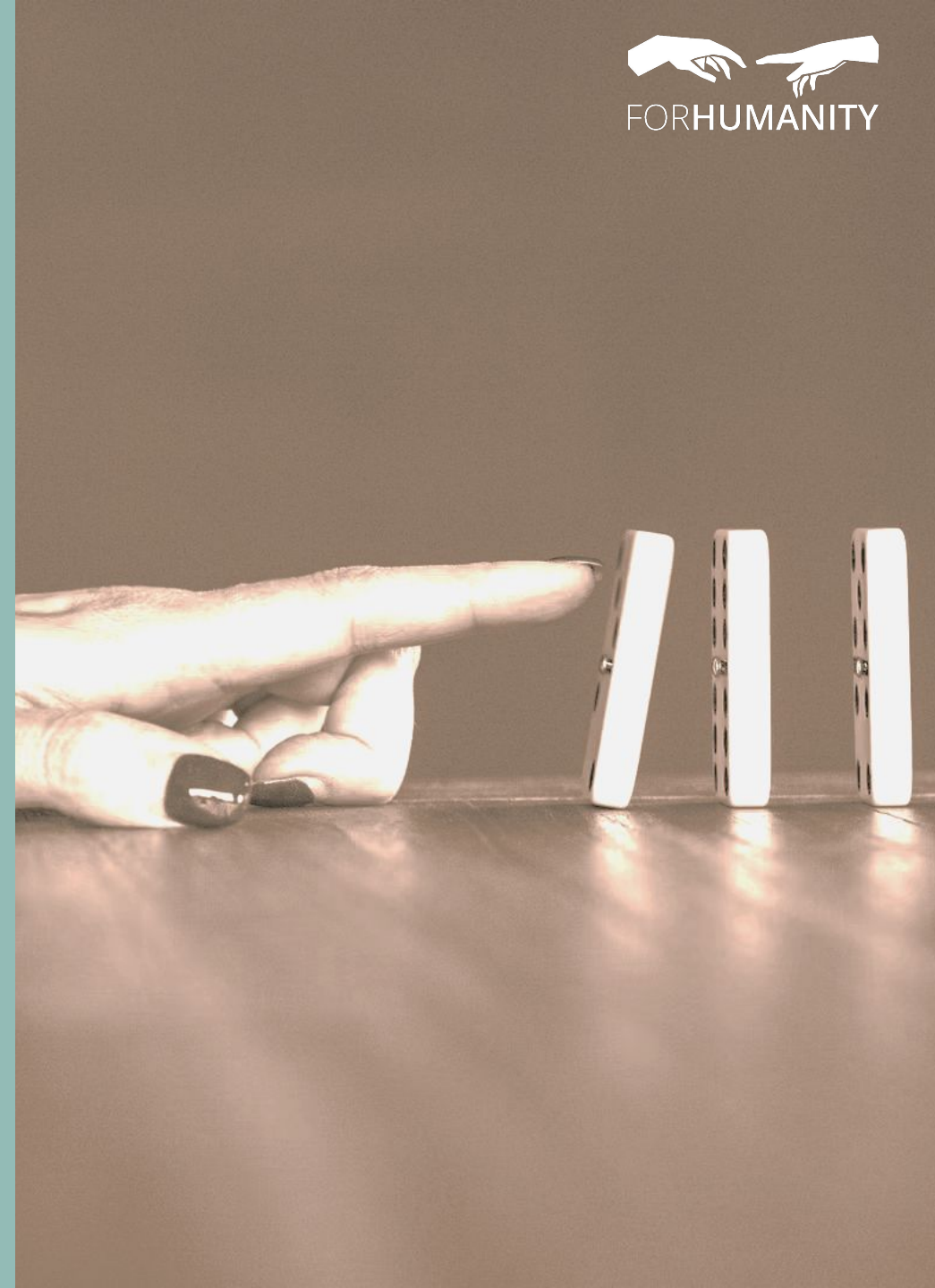


Ask Yourself: What are the Consequences?

Who can be negatively impacted by using this tool?

- Me
- Employees
- Shareholders
- Customers
- Management
- Environment
- Society/Community
- Corporate principles:
 - Brand
 - Trust
 - Customer base
 - Loyalty
 - ESG

TEST: If the benefits outweigh the risks then it seems reasonable to continue and you have overcome one aspect of automation bias



Education and Training

Focus first on giving the Human in Command what they need

For Human in Command - on the AAA system

- Typically vendor training for IT
- But nothing for many AAA Systems
- Substitutes:
 - Users' guide (if available)
 - Model cards (if available)
 - Understand the inputs
 - Understand the explainability of the outputs
 - Third party?

For human in command of the Human in Command

- Understand the higher level concerns
- Identify the gaps
- Support the HiC with what they need



7. Taking Action Against Automation Bias

Overcoming automation bias is an active process. It requires designers, developers, deployers and users to actively consider how automation bias might be affecting their process or evaluation in working with a AAA System



Provide Appropriate Human Oversight

Oversight teams should assess the AAA System to determine the appropriate level of human involvement



1. **Human-in-the-loop** (any model that cannot offer an answer or conclude processing without human intervention)
2. **Human-on-the-loop** (human supervision and/or control but the system is able to conclude processing without the need for human intervention)
3. **Human-in-Command** (as described by the EU AI Act, a natural person assigned by a system Provider or Deployer to act as no less than a third line of defence governance, knowing the capacity and limitations of the system, possessing sufficient training for the regular operation including the identification of anomalies, dysfunctions and unexpected performance)
4. **Post-hoc human review** (in the context of the scope, nature, context, and purpose of an AAA System, an assessment conducted by a team of expert reviewers designed to determine if the model remains healthy and fit for purpose, including performance assessment against a set of metrics, measurements, and thresholds)

Steps to Overcome Automation Bias



Education and Training

Users of AAA need to recognize the potential for automation bias. Encourage critical thinking and vigilance



Increase Accountability

Making users accountable for either their overall performance or their decision accuracy leads to lower rates of automation bias



Encourage Human Input

Design interfaces that facilitate human input and collaboration, ensuring that users remain engaged in the decision-making process

Approach the Problem With an Open Mind

Avoid Preconceived Notions of Causality



Require Checks on Automated Tools

We need to establish monitoring and controls on AAA Systems - to ensure model health and fitness and stay true to its purpose



Build Alert and Verification Systems

If we doubt everything we hear, read, see we cannot avoid the rabbit-hole of paranoia - so instead build tools to monitor performance



Combat Your Own Inherent Biases

Seek diverse input and multi stakeholder feedback to ensure different perspectives and ways of thinking to assess the risk of bias

UI Mechanisms to Combat Automation Bias

Profiling Decline



User can opt out of recommendation engines or other content moderation systems that require profiling the user

Profile Reset



User can zero-out or completely reset the profile created by the provider of the system for the users interface with the AAA System

Profile Re-engage



User can re-apply their profile to the AAA System after a period of profiling decline or profiling reset



Providing Feedback to Improve the System

Find ways that you can help



- Use employee systems to raise questions, and concerns (e.g., illegal use of IP)
- Use Adverse Incident Reporting Systems
- Report incidents to public domain collectives, such as AI Incident Database



Why and How Do We Trust?

It takes an ongoing, continuous process to “avoid” Automation Bias

Everything you have just learned...

Be prepared to do it again tomorrow

and the next day

and the next day



Overcoming Automation Bias

Thank you

- and good luck!



FORHUMANITY

