FORHUMANITY

COMBATTING AUTOMATION BIAS

CAN AI UNDERSTAND THAT I DON'T WANT TO BE MY HISTORY?

February 2023 | **https://forhumanity.center/**

Source: Photo by Brett Jordan on Unsplash

This article will guide the reader through the context of automation bias and how it discriminates people from having an ethical choice to be not their history. It examines specific examples and philosophical context to provide a view of the need for humane efforts to make AI respect humans and humanity (rights, freedoms, equality, dignity)

TABLE OF CONTENTS

INTRODUCTION

Artificial Intelligence and Algorithmic Systems continue to proliferate. Enamored of efficiency, statistics, and sheer power of processing, corporations, designers, developers, and now end users (ChatGpt) continues at a breakneck pace to deploy these systems into increasing human tasks and the result is that their impact on individuals is increasing. As these tools proliferate, the need is rising to combat Automation Bias (defined as "*occurring when a human decision maker disregards or does not search for contradictory information in light of a computer-generated solution which is accepted as correct*"), especially as the users of these systems are increasingly amateur operators who are oblivious to the risks, inaccuracy, and unintended consequences associated with these tools. Left unchecked and without robust mitigations, the AI hype train increases Automation Bias and endangers three inherently precious human values: redemption, forgiveness, and the right/ability to change.

Photo by Engin Akyurt

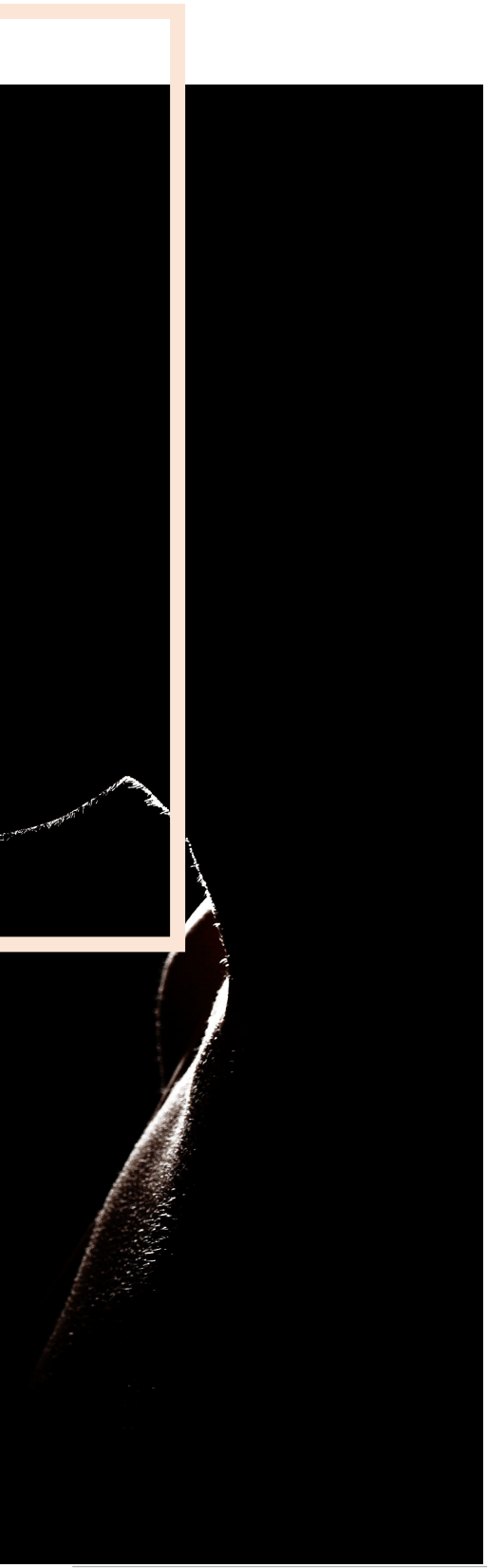
TURNING POINTS

Redemption, forgiveness, and change have one specific thing in common - they are turning points. Let's anchor our discussion in definitions:

- **Redemption** is *the act of making something better or more acceptable.*

- **Forgiveness** is *a conscious, deliberate decision to release feelings of resentment or vengeance toward a person or group who has harmed you, regardless of whether they deserve it.*

- **Change**, while relatively well understood, let's contextualize its use in this paper as - *to make radically different, transform.*



Going a Little Deeper…

Redemption – countless real-life stories are often depicted in Hollywood movies celebrating redemption. The hero poised for great success has a setback, gets off track, looks spiraling out of control to ultimate failure, and yet a turning point occurs. The path is reversed, and, as the story goes, victory is attained through hard work, blood, sweat, and tears (often courtesy of a musical montage). There is great appeal to our humanity in such stories and, thus, the oft-repeated theme.

Social approaches have often found methods to enable redemption, including alternative punishments, atonement, and community service. Such redemption gets reinforced with good behavior over time and gets naturally accepted by the community by slowly erasing the memories of past misdeeds.

Redemption stories stir empathy and hope in each of us. We can relate to challenging times, downtrodden episodes, and difficult seasons from our own lives, and we hold onto the hope of redemption and scan the horizon for a turning point - for some, it may be the *only* positive to cling to as we face each day. Hope fosters will drive and energy. It supports innovation and problem-solving, leading to humanity's turning points.

Forgiveness - considered by many as a dying art, and yet critical for our health, whether physically, as discussed by Johns Hopkins,

Studies have found that forgiveness can reap huge rewards for your health, lowering the risk of heart attack; improving cholesterol levels and sleep; and reducing pain, blood pressure, and levels of anxiety, depression, and stress. And research points to an increase in the forgiveness-health connection as you age.

Or, from a more philosophical perspective, covered in literally thousands of texts dating back thousands of years, captured by Rabbi Jonathan Sacks

[3] Forgiveness: Your Health Depends on It - Link
[4] A Brief History of Forgiveness, Rabbi Jonathan Sacks - Link

It is one of the most radical ideas ever to have been introduced into the moral imagination of humankind. Forgiveness is an action, not a reaction. It breaks the cycle of stimulus-response, harm and retaliation, wrong and revenge. It frees individuals from the burden of their past, and humanity from the irreversibility of history. It tells us that enemies can become friends.

From either perspective, forgiveness's great value to humanity is clear. A release, freedom, again - a turning point. Many have argued that forgiveness is one of the greatest expressions of power and love, an idea expressed eloquently by Dr. Martin Luther King:

We must develop and maintain the capacity to forgive. He who is devoid of the power to forgive is devoid of the power to love. There is some good in the worst of us and some evil in the best of us. When we discover this, we are less prone to hate our enemies.

"We must develop and maintain the capacity to forgive. He who is devoid of the power to forgive is devoid of the power to love. "

Right/Ability to change - so fundamental to our nature as humans that it isn't even listed in the 30 articles of the UN's Declaration on Human Rights. However, if I told anyone reading this article that they were not allowed to change, you would be deeply offended and assert the exact opposite. The explanation for the explicit omission can be found in the fact that a "right to change " is inherent/implied in almost every one of the 30 UN Declaration on Human Rights Articles (it could even be called the common thread). The right to change finds impetus from a person's negative situation and innate desire to seek remedy. Freedom from slavery, Freedom from being uneducated, Freedom to move, and on and on through the 30 articles. Every article is an expression of a negative, sub-human level of treatment and collectively, we say, "no" that is unacceptable. Every person has the right not to be treated that way - an exact right/ability to change one's circumstance.

Extrapolating the right to change out further, there are no earthly-bound pre-destinations. There are no pathways that cannot be strayed from, there are no endless journeys devoid of an opportunity to deviate. Our very humanity *needs* the right/ability to change and more importantly, this right/ability should be celebrated, protected, embraced, and held beyond reproach.

Incompatibility with AI and Algorithmic Systems

Photo by cottonbro studio

Redemption, Forgiveness, and our Right/Ability to change are aspects of our humanity to be treasured. Turning points are rooted in hope and precious to our humanity. However, here, we generally encounter a problem with AI and algorithmic systems. They don't *do* turning points.

As a result, we already have applications of AI and algorithmic systems resulting in a detrimental impact to people with no such opportunity to seek forgiveness or redemption. Some instances include:

1. Algorithmic systems predict someone to be an offender based on past offense history;
2. Algorithms rejecting candidates based on their scores in school;
3. AI-driven background check tools providing red flag report based on past behavior (including social media behavior);
4. An algorithmic student grading system that graded students based on their past grades, specifically during COVID-19

Artificial intelligence and algorithmic systems do not share these human values and, based on programming, ignore, repress, and actively fight against them. 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. The very nature of AI and algorithmic systems causes them to continue on course - without deviation.

Deepening concern occurs when we realize that the more data, the more lived experience, the deeper the trendline, then AI and algorithms will infer the higher likelihood of predictive power. These inferences are treated with greater voracity and expectations move from an inference to a "fact" (another form of Automation Bias). Imagine a downward-sloping trendline measuring a person's depression levels. These moments cry out for a turning point, and not another step towards darkness, yet using today's modeling that turning point is rendered impossible. They "predict" that you cannot deviate from your history.

Our Cognitive Bias (Automation Bias) Used Against Us

The aforementioned “fact” emerges from a collective cognitive bias that inferences deduced from an AI, or algorithmic system have predictive value. Models such as these are built on likelihood but they do not express causation. Correlation, the hallmark of AI and Algorithmic systems, does not equal causation and cannot determine with absolute certainty, or even beyond a reasonable doubt, what the next occurrence will be.

Consider validity, accuracy, and associated risk from AI/ML model validations. Such models may advise that they have been tested to be 80% accurate or 95% accurate. Coincidentally, in your mind, you hear and think accordingly, "20% or 5% inaccurate". You know there is a "reasonable doubt" associated with the model, enough that you will question and critique the conclusion, largely because you have been provided with the information. This informed thinking rightly infects your acceptance of truth or fact regarding the model. Associated conclusions are accepted and understood to bear risk - the risk that they are wrong/incorrect. These balanced and risk-based thought processes are good, as we will discuss below.

However, somewhere along the way for each of us (often context-dependent) - as the tested accuracy rises to 99% or 99.99% or 99.9999999999% - our mind changes and begins to interpret this model to be rendered "truth" or "fact" and our reasonable doubt wanes or even disappears. Our internal risk management mechanisms cannot properly value 1%, .01% or .000000001% chance as being meaningful, likely, or even a risk worth trying to manage. This natural occurrence is deeply dangerous to redemption, forgiveness and the right/ability to change and it represents another form of Automation Bias.

The Impact of Statistics on Our Own Decision Processes

We describe this phenomenon as *neglect of probability bias*, a cognitive bias, not dissimilar to the way our minds bend and alter absolute truth, like in the gambler's fallacy.

Photo by Pixabay

The gambler's fallacy highlights our tendency to conclude that if 5 or 7 or 10 flips of a coin render HEADS, then we believe the subsequent likelihood of TAILS increases for the next flip - when, in fact, the probability of a head or tail never changes. When we assess the future, our cognitive bias changes our expectation from the ground truth of the event we are trying to measure to an assigned, false expectation, assuming the past has meaning for the next event.

Returning to the accuracy of our model, the *neglect of probability bias* alters our evaluation of the model to conclude that the model is so accurate - its decisions must be true or that the factual probability has changed because of past occurrences. Some projects even further on these tools and assume truth because it is a machine. Under those assumptions, there is no room in those conclusions for redemption, forgiveness and the right/ability to change. The probability, in our minds and based on history, leaves no reasonably acceptable chance for such a turning point. Now, our history becomes encoded for our future and we are preventing, hindering, and outright rejecting the opportunity for change. To be human and celebrate the best we have in our collective spirit, we cannot accept this fate and must insist upon better embedded ethics in AI and algorithmic systems. Choosing to reject this fate begins with combatting Automation Bias.

The conclusions reached by all of these systems is NEVER a fact. They are *always* an inference, and as the General Data Protection Regulation (GDPR) has codified - inferences are not facts. The inferences used as predictions, determined by our AI and algorithmic systems are not fact - because they cannot measure the independent probability of the next decision. Our past is not the only input to future decisions. Each new decision might be our time for redemption, forgiveness, and most importantly, our turning point, simply because we have that right. Said differently, there are two realities. The first is the one just concluded by the model - based upon the past - "here is what is expected to happen," the model concludes. The second ought to be an equally meritorious model of "what is the independent likelihood of change for the next decision". This second model isn't being built today nor applied to our recommendation engines, content moderation, news filters, and systems designed to adjudicate our futures, like creditworthiness or criminal sentencing algorithms.

Why does it matter?

It matters because there are companies, firms, designers, and developers working to exploit these cognitive biases and have the inferences they deduce from AI and algorithmic systems be interpreted as "facts" so that they can be relied upon by decisions makers, leveraged into conclusions about you, or directly into goods and services offered to you - preying on Automation Bias. This is an industry worth more than $200 billion annually just in advertising based on these "facts". When this occurs, you have been stereotyped by the model and by the organization operating the model.

As the brilliant paper from Anita Bernstein in 2013 What's Wrong with Stereotyping stated, "A*nyone who asks what's wrong with stereotyping*[17] *accordingly, has to stipulate that not all stereotypes are wrong in the sense of doing harm that is severe enough to warrant sanction from the law, a costly response. Some stereotypes might be false or unreliable but do not offend the groups of people they reference.*[18] *Some might offend but have the virtue of being reliable, or true enough."*

Stereotyping is considered wrong in the legal sense, Bernstein continues when she says, "*The answer... is that stereotyping is wrong to the extent that it functions to deprive individuals of their freedom without good cause."* The reason this matters to you is that humans have a right to "not be our history", to not succumb to or abide by assigned stereotypes, but instead to earn redemption, to offer forgiveness. Ignoring redemption and forgiveness in our AI and algorithmic systems will withhold these evolutionary concepts that valued humanity, increased harmony, upheld dignity and resulted in stronger communities.

Photo by Mizuno K

Photo from Microsoft

Coding and Embedding Human Ethics in our decision-making tools

Redemption, forgiveness, and the right/ability to change are too valuable to humans. They are human values to be emulated and encouraged; therefore, if we live in a world filled with socio-technical 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.

AI and algorithmic systems limit choice through recommendation engines, content moderation, or by assuming a continuation of the previous behavior and systematically rendering a decision on past information. The process being modeled is reasonable and, unfortunately, simplistic. Deducing that a person will make the same decision tomorrow based upon those previous 999 instances has merit. However, there is a distinct difference between a guess, an estimation or even a likelihood versus a pronouncement or an inference from a machine that limits choices, and information and forces the next conclusion upon the individual. Unfortunately, the systems we use today are designed to do exactly that.

AI and algorithmic systems are deployed to predict truthfulness, emotions, creditworthiness, future health, or even the likeliness that you have mal-intent. As a society, we cannot accept a future determined only by our past and must recognize that there can be two correct and reasonable states of likelihood, one based on the past and a second that treats the next occurrence as a fresh opportunity, especially if it is a hopeful change or in our collective best interest. These moments of change, of forgiveness, of redemption should be rewarded, encouraged, and embedded in our socio-technical systems. Some will argue that there cannot be two *probabilities* for the next event, but our brains and thought processes accept this dichotomy frequently. This two-mindedness about the likeliness of a future event forces us to think critically about reconciliation, a step vital to our humanity and overcoming Automation Bias.

Critical Thinking

Critical thinking, as defined by the American Philosophical Association says - *the process of purposeful, self-regulatory judgment. The process gives reasoned consideration to evidence, contexts, conceptualizations, methods, and criteria.* Overcoming Automation Bias starts with critical thinking and a healthy skepticism regarding our AI and Algorithmic tools. Recognizing that the decisions they offer are one set of solutions, not to be blindly except, but considered, evaluated, and analyzed.

Our continuing example can demonstrate an example of critical thinking. **Assume a person chooses to eat Cheerios 999x in a row. If only that historical data is what is fed into our AI or Algorithmic system to predict our future breakfast choices, it is easy to anticipate the prediction. Under these conditions, our AI and Algorithmic systems will conclude with a high degree of certainty that we will eat Cheerios tomorrow.**

Photo by Eric W.

Now, we can examine a range of exogenous, non-historical variables that might impact the next decision -

1. The person is not in their home for the next breakfast
2. They did not bring breakfast with them
3. They are camping in the woods tomorrow morning

Let me ask you, now that you are aware of those facts, do you still think the chance of the person eating Cheerios is 99.9%?

To return to the analogy of *neglect of probability bias* and gambler's fallacy. In this instance the probability of a new decision is not .001%, it is in fact, substantially higher and was a higher probability for every one of the previous 999 instances. AI and algorithmic models have no choice but to infer the statistical likelihood - it is in their coding. A further manifestation of Automation Bias is that the designers, developers, and data scientists deploying these tools often believe that they have "correctly" modeled your behavior - that they understand you. Critical thinking allows us to understand and appreciate that tomorrow, Cheerios might be in the bowl, but if they are not, we won't be shocked by a 1 in 1000 event occurring because we have the ability to understand the independent probability of the next decision (historical plus non-historical data inputs are relevant).

We have used this word "independence" (statistically independent) a couple of times and it should be defined properly as: *the occurrence of one event does not affect the probability of occurrence in another event*. Continuing with our coin, the probability of a head or a tail is always 50/50. That statistical independence is true regardless of the previous collection of coin flips and can be extrapolated to being critical to our humanity, because in independence we find redemption, forgiveness, the right/ability to change. It is where we have the right to "not be our history".

That statistical independence is true for humanity because, in independence, we find redemption, forgiveness, and the right/ability to change. It is where we have the right to "not be our history".

Considering the truly independent nature of each decision and recognizing that the past is far from a complete set of relevant variables, now we know why critical thinking about the decisions rendered by AI and Algorithmic systems, like ChatGpt, is necessary. Understanding when tools will be robust and when they *must* be questioned highlights the importance of critical thinking to overcome Automation Bias. Critical thinking allows the tool to remain useful but demands that the user recognize that a historically-oriented tool isn't always right or accurate because the next decision is often truly independent and far more complex than just "what happened in the past".

Over reliance on historical data can lead to some absurd results. Something many of us have experienced. Imagine you are shopping for a new pair of earbuds online, doing search queries for the best pair for your needs and within the first hour, you hopped onto the company website and bought the earbuds. Unfortunately, in our world, where data is captured and resold, your interest in earbuds is flying from search engines to data brokers to advertisers, in the name of profit and the oft-justified, but misguided, belief that you will benefit from receiving advertisements for earbuds. So now, earbud ads follow your every movement online. YouTube, Amazon, CNN.com, CNBC.com, all your favorite websites for weeks. This is an example of reliance on "your history determines who you are". Advertisers are paying money for the information that you were searching for earbuds, but that was "you" for an hour and now "you" are someone else - someone that certainly has no need for earbud commercials. Advertisers lose too, they really needed to know that you made your purchase and an earbud shopper no longer defines you. Both you and the earbud advertisers lose, because of the way our models are developed and an anchoring to historical data.

Consequences due to automation bias

Anchoring to the past breeds hopelessness or as George Santayana quipped, "Those Who Do Not Learn History Are Doomed To Repeat It." - another version of hopefulness. "Learn" is the keyword in Santayana's wisdom for overcoming the negative outcome - repeating history - hopelessness. "Learning" is one way to enable our ability to change. Failure to learn from our history or more precisely to the aforementioned example, to over rely upon our history in our AI and Algorithmic systems may lead to a negative spiral devoid of change and personal growth.

It is our hope that humanity broadly chooses to embrace the best aspects of our humanity and avoid Automation Bias. This choice is not the easiest path, but it is ForHumanity's belief that it is the path that leads to hopeful outcomes that seek redemption, offer forgiveness and always ensure that we have the right to change leading to greater well-being for us all. Wouldn't we all like to see more redemption, more forgiveness, and AI and Algorithmic systems that uphold and support our right "not to be our history" and therefore embrace opportunities for change.

Photo from Microsoft

TOOLS TO EMPOWER CRITICAL THINKING AND COMBAT AUTOMATION BIAS

Below we offer some tools/procedures/processes that will actively combat automation bias. We will embed these in ForHumanity Independent Audit of AI Systems certification schemes. Still, we hope that every person who designs, develops and deploys these tools increase their awareness about the risks of automation bias and endeavors to overcome it.

We can improve and increase information, heighten awareness, and the need for critical thought as we deploy AI and Algorithmic systems. Also, we can demand models that include tools and nudges for our best interests. These suggestions encourage and operationalize overcoming automation bias:

1. Transparency
2. Disclosure,
3. Pause buttons,
4. Demanding support for the user's best interest instead only profit maximization (Nudge Techniques)
5. Training and Residual Risk disclosures

All of these tools are available to us but need to be addressed or even at odds with creating shareholder wealth (a primary goal of the providers of these technologies).

Transparency

Transparency applied to our AI and algorithmic systems is a valuable first step. In the recently passed Digital Services Act, there is a requirement for Covered Entities using Recommender Systems to conspicuously declare their usage and share with the recipient of the service an explanation of "how and why" those recommendations are being offered.

Disclosure

Disclosure, a higher form of transparency that requires documentation and details associated with the aforementioned transparency. Disclosure might include the explanation of prior content moderation or recommendations applied to you. Let's take Facebook/Meta as an example, specifically your posts and news feed is filtered. It is designed to show you posts you are more likely to engage with, to respond to, and generally interact with. Now imagine if you were met with an honest disclosure that read "we have been providing you with content recommended for your tastes and designed to increase your likelihood of response". Like the transparency example, an explanation like this will cause a defensive reaction in people. Regardless of whether we can require companies to provide these disclosures, we need to train people to operate with this type of defensive posture when they are spoon fed their digital services.

Residual Risk is a critical tool for ensuring that AI and Algorithmic system users are "informed users". An example of a Residual Risk disclosure is the warnings the FDA requires for drug makers selling pharmaceuticals. Commercials about drugs are required to state their benefits and immediately follow with a list of side effects as evidenced in clinical trials. With this information, a person is now equipped to understand the possible benefits and the potential risks of taking this drug. Informed users are less likely to suffer from Automation Bias.

Photo by Anna Nekrashevich

Pause and Pause Buttons

Continuing with the Meta/Facebook example, a pause buttons designed to actively stop the AI or algorithm content moderation/recommender system from providing you with a tailor-made filter. This pause impacts us in three valuable ways:

1. it provides time for a person to think about the perspective they have been seeing
2. allows the person to step outside of the system and see an unfiltered view in order to reassess the "value" of the content moderation, or recommendation choices.
3. We are reminded of alternative choices and perspectives. It reminds us that there are often two or more sides to a story or argument.

4. Most importantly, a "Pause" provides a specific and deliberate opportunity for change.

Pauses in the systems and tools we use are a good way to be reminded that filters are by definition limiting. They remind us to engage in critical thinking.

Nudges

"Pause provides a specific and deliberate opportunity for change. Pauses in the systems and tools we use are a good way to be reminded of critical thinking."

Nudge techniques have become common, if not ubiquitous, in our online experiences. Highlighted by the movie Social Dilemma, the UK's Children's Code, California's Age-Appropriate Design Code, and significant academic work, online service providers have been pushing us in directions that increase their profitability. While these nudges might be in your best interest, they are equally likely to be against your best interest depending upon the business model. In either case, they are taking advantage of your innate preferences and guiding you where they want you to go. So another tool to combat Automation Bias and where we could provide nudges would be towards redemption, forgiveness, and the right/ability to change. Nudges designed to foster and model the turning points always available in our lives.

To examine how nudges work, we can continue with the Cheerios analogy, measuring all past choices and achieving a high likelihood of eating Cheerios tomorrow. Next, we can introduce one of the phrases below as a pre-breakfast nudge (easy enough to code):

1. "FYI, you've eaten Cheerios for 999 straight times - you can't eat Cheerios today."
2. "You MUST eat Cheerios today"
3. "Oatmeal is more healthy than Cheerios"

In the UK, we have begun to see best-interest, nudge requirements demanded by law. One example can be found in the gold standard law - the Children's Code. Here, the law demands that the rights of the Child (as identified by the UN's Convention on the Rights of the Child) must be treated "on par" with shareholder value. A difficult and challenging standard to achieve governance, accountability and oversight upon, but one that ensures the choices offered to a Child regarding their Personal Data is one that is predicated on the Child's best-interest and not shareholder interests alone.

Training

Lastly, ForHumanity will be offering training for people on critical thinking and combatting Automation Bias. Our training is designed to heighten awareness, provoke critical thinking, and equip all users with the necessary tools to overcome Automation Bias. When we remember that AI and Algorithmic systems are tools to empower and support humanity then we will ensure the best possible results from these tools…ForHumanity.

Photo from Microsoft

FORHUMANITY

Photo by Javardh on Unsplash