IS UNBIASED AI POSSIBLE?

infoshare | Sagacity

We've long discussed how companies relying on unverified data for decision-making have suffered reputational damage and caused unwarranted harm to both existing and prospective customers.

There have unfortunately been some well-documented examples: public sector bodies who have sent personal information to the wrong addresses; threatening letters sent to homeowners based on the wrong information on mortgage arrears; incorrect levels of risk and exposure... the list goes on. We need industry-wide acceptance that successful AI relies upon a solid data foundation, coupled with suitable scrutiny and challenge by humans to prevent bias from creeping into automated outcomes.

A recurring theme debated at our recent 'Eliminating bias in AI models' workshop was knowing the right questions to ask of the AI to determine bias. If human questions spot unhealthy bias, can the algorithms be easily changed to prevent similar future biases, or is a 'scrap and rework' programme required? When dealing with sensitive data or data concerning vulnerable individuals, I believe a higher level of scrutiny is needed, compared to data related to assets, aggregated profiles, or non-critical mission environments.

Pamela Cook
CEO, Infoshare

For years, AI has been a staple of sci-fi movies. One of my favourites is Minority Report, in which personalised advertising and intelligent whiteboards were introduced as science fiction: they have since become a reality.

Looking ahead, the future of crime detection lies in prevention, a concept that can be powered by AI. Imagine a scenario where AI predictions, rather than a woman in a bath, form the basis of crime prevention strategies. This is not a far-fetched idea, considering the advancements in AI technology. As AI becomes more prevalent, the idea of using past data to predict future crimes doesn't seem so ridiculous. Data is already used to provide background information on locations being visited and external support requirements, such as translation. Using similar data to predict crime hotspots is undoubtedly already being investigated.

However, it opens a can of worms: what if our historical data is not representative? As the make-up of our population changes, how do we ensure that the predictions we create are not biased by what has happened in the past? It was enjoyable to debate this at our workshop, and to share the outcomes in this eBook.

Scott Logie
CCO, Sagacity

Introduction

Can AI remove the human biases that reinforce inequality, or will using it perpetuate harmful ideologies?

Alongside our partner Sagacity, we recently ran a session at the Women in Data flagship event, about eliminating bias in AI models.

We hoped the topic would resonate. But when it sold out within 10 minutes, we realised that it was a more recognised concern than we thought.

After the initial wave of excitement about the possibilities of AI, the conversation is shifting to how it can be used safely, legally, and ethically. Bias can creep into data, for instance, through flawed data collection methods or inherent societal biases. When biased data is used to train AI models, those biases can become embedded within the AI, delivering outputs that reinforce prejudices. If these outputs are used to inform decision-making, harmful ideologies can be perpetuated instead of eliminated.

At the workshop, we covered the topic from three angles:

1. Can we ever really eliminate bias?
2. Should all biases be eliminated, or are there circumstances where they're required?
3. How do we eliminate bias from AI models, and how can we monitor and regulate usage of AI models?

1.Can we eliminate bias in AI?

All data is naturally backwards looking; it documents what has come before. Making decisions based on the past might not lead us to the future we want to see and may just give us more of the same. As Henry Ford famously said: “If you always do what you’ve always done, you’ll always get what you’ve always got.” If data is used to train AI without appropriate consideration of the potential inherent issues, it can impact the reliability of any insights we draw. Although eliminating bias might be an unrealistic goal, recognising where it’s coming from is crucial to reducing its impact.

Before setting up AI models, these were some of the main considerations and mitigation strategies discussed at the session:

Understanding data quality

If we don’t understand the quality of the data going in, how can we rely on the quality of the insights coming out? From data collection methods that aren’t representative to data input errors that skew results; understanding where and how data quality can degrade can help to frame the resulting data insights and signpost areas of weakness in the process that need to be targeted to reduce bias.

Therefore, it’s vital to identify and address errors, inconsistencies, and missing values to improve data quality.

Human-in-the-loop decision making

AI and automation will change the landscape over the next 10 years, yet we must be careful not to over-rely on technology.

For example, it can be used to replace human decision-making in areas where unconscious human bias has the potential to lead to discrimination, such as in employment hiring processes. However, if a ‘good candidate’ is identified using data on previous hiring decisions, any past discrimination becomes baked into the AI model.

People will always be needed to sense check AI decisions and to examine whether prejudiced patterns are forming in the AI outputs.

Accepting knowledge limits

One thing will always be true about data - we don't know what we don't know. We should always be aiming to understand the data, and how we should examine the areas of potential bias. However, we must accept that at a given time, we might not have all the information available to us to make accurate decisions, and there may be bias in the data that we're not aware of.

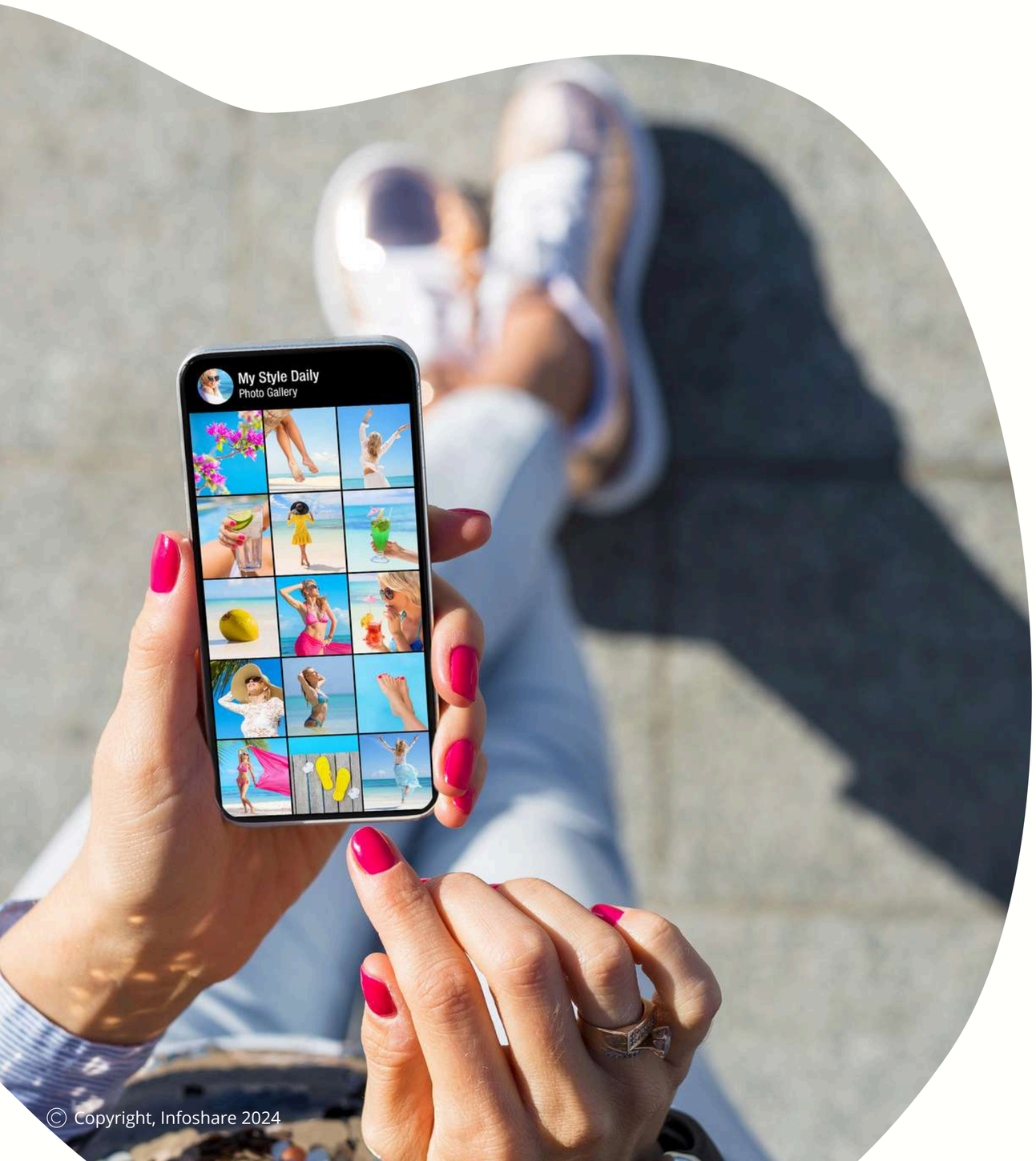
Continual re-evaluation is required, along with a willingness to rethink previous insights if new bias is identified.

Improving data transparency

Having transparency of our data and where it comes from can help to identify and evaluate any inherent biases. Achieving full transparency over a data landscape is a challenge. But it's only once we're confident that we have full visibility that we can understand what bias exists, where it comes from, and most importantly, what to do with it. Not all bias will necessarily be harmful, and not all bias can be removed. It needs the human element to decide what to allow, what to mitigate, which questions to ask, and what to eliminate.

2. Should **all bias** be eliminated?

This brings us to the second topic we explored at the workshop – should we be aiming to eliminate all bias from AI models? Are there circumstances where they are unavoidable, helpful, or have the potential to act as a force for change?



Depends on what it's used for

There are occasions when we want to understand the data in its true form, even in the knowledge that it contains bias. For instance, there are circumstances where it is important to maintain historical accuracy, even if the outputs aren't representative of the world we see around us today.

Also, bias in models can be useful. A good example is social media or streaming service algorithms that filter content based on the age of users, to prevent inappropriate content being shown to younger users.

Eliminate or mitigate negative bias

There was general agreement that, as far as we can, we should look to eliminate negative biases – that is, those that reinforce inequality or harmful stereotypes. But not that we should eliminate bias completely, just negative bias, where it can be detected.

This obviously relies on the ability to both identify the bias and find a way to remove it. In terms of identifying bias, it's important to have a diverse team responsible for validating the data feeding AI models and checking the outputs for signs of bias. But when we can't eliminate, we should look to mitigate the impact, or at least caveat the results and ensure we don't overextend our assumptions.

In this example, the bias should be acknowledged, as well as the potential negative impact of extrapolating the results across unrepresented demographics. And when using the insights, care should be taken when making assumptions about a group much wider than the original research sample.

Take the example of using external data. When using third party market research, we find there is a bias in the data collection method (such as people from a particular socio-demographic group overrepresented in the participant sample). We can look to eliminate this bias, by searching for more representative research or commissioning our own. But if there isn't more representative research available, or there isn't the budget for commissioning our own, what action should be taken?

Making it up - synthetic data

Synthetic, or imputed, data can be used to fill in the blanks or create more representative samples. There has always been a problem in statistical modelling for rare diseases where the sample of those who have the disease far outweighs those who do not. Sampling can create an unbiased dataset but that can leave us with a low volume of observations to work with. In this case, imputing or creating false data from the real data provided is not an unusual occurrence, and the same approach could, and should, be considered for building larger datasets to prompt AI models.

For example, in a healthcare scenario where not enough historical data has been collected for some conditions in particular ethnic groups, synthetic data is often used to ensure that sufficient combinations are input into the models to help differentiate diagnosis.

However, a concern was raised that synthetic data could simply change the problem into a new one, replacing the original bias with a new one. Therefore, care needs to be taken with this approach, and thorough testing to check for any new biases that emerge.

Recognising bias can highlight the need for change

Removing bias or balancing it with synthetic data might change the effect it has on a particular AI model. But what else? During our discussions, the point was raised that if we only solve the impact of the bias on our own model, we miss an opportunity to highlight the need for change in the broader societal context.

A real word example mentioned at the session was the lack of data on women's injury in sport. Removing the male data, because it's only relating to one sex category, is not the answer, as research into men's sporting injuries is still important. Similarly, creating synthetic female data based on the male data could be dangerous, as there will be different physiological, contextual, and societal reasons behind the injuries between men and women.

The answer then, is to use the fact that there is bias within the data to advocate for change within industry, government, and society as a whole; we should take opportunities to address bias in the real world, not just in our AI models.

Conscious positive bias as an agent for change

When we recognise there are inherent biases at play in particular contexts, even if we can't eliminate them, we can build positive bias into our AI models to counteract their impact. For instance, how could a company use conscious positive bias in hiring decisions? If they were aware that, historically, women have faced discrimination in hiring decisions and, as a result, men have been disproportionately more successful in their applications, they could build in conscious positive bias to an AI model to counteract the impact for future hires.

Even if negative bias isn't already present within a dataset, AI models can still be consciously built with positive bias when we're looking to use the models to have a positive impact on society. For people and organisations looking to provide opportunities to improve equity for underrepresented groups, conscious positive bias can be used for good.

But always remain aware

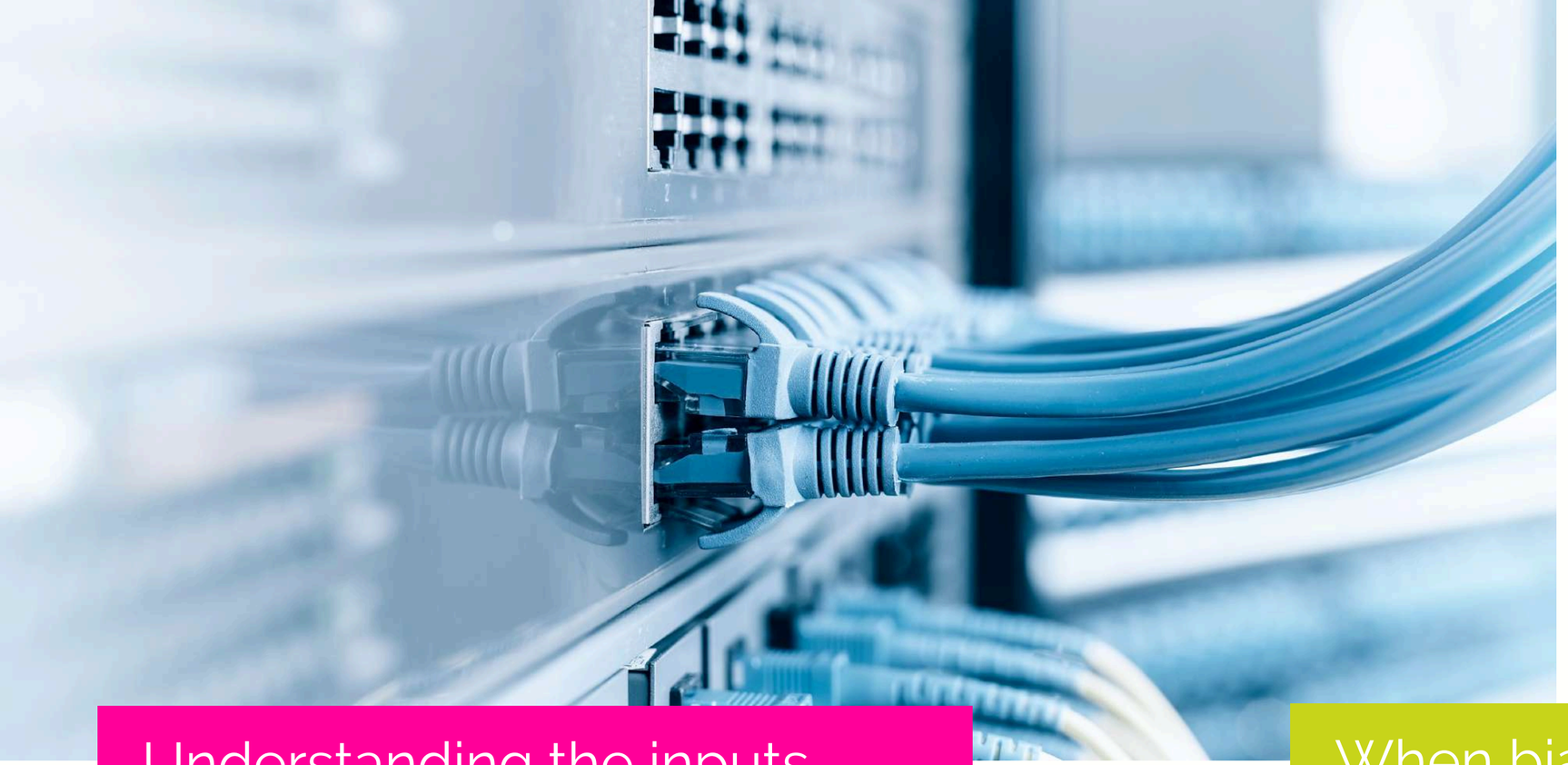
The hardest aspect of this challenge is understanding what bias has gone into the models created. There is work to be done upfront to look at the underlying data being used to prompt the models, and to ensure that data is unbiased. But there is also work to be done in validating the model outputs to ensure that the scores, decisions, and results created are also not biased (unless that was a goal, of course).

One suggestion is to ensure all models are validated and checked by a diverse group of people to ensure unconscious biases don't get missed by those who may not be able to recognise them.

3. How do we deal with bias?

Understanding discrimination and bias pitfalls

A full understanding of the different types of discrimination, and how these can present in both overt and subtle ways, is important for understanding whether AI will perpetuate discrimination with its outputs. More diverse and inclusive organisations will naturally benefit from a diversity of thought and experience to create detailed conditions for the data if it is to be considered an appropriate feed for AI models.



Understanding the inputs

To promote equality of outputs in our AI models, we'll need to be alive to potential inequality in the inputs. To do this, we need a proper understanding of the data we're using – and not using – as our inputs. Are there any gaps or blind spots in the data that could have unintended consequences?

When bias is identified

As mentioned earlier, if we do identify bias that, through proper assessment, we believe will have a negative impact on the model and influence the outputs, we must seek to eliminate the bias where we can. This can be done by:

- Removing the data feed completely
- Recognising the limitations of the data and tightly restricting its use
- Balancing it with additional positively biased datasets aimed at improving equity

Implement tight controls and regulations

AI alone cannot solve our problems with discrimination, and it needs to be used in a regulated and controlled way. KPIs and controls should be set by a steering group who agree ‘what good looks like’ and ‘what bad looks like’.

The data being used, and the models themselves, should be regularly reviewed and assessed by a team responsible for monitoring their levels of bias. This will ensure that companies are transparent and accountable for the bias in their data, and supports an overall more ethical approach to the use of AI.

Infoshare is a UK data quality and data management company with a mission to give organisations the confidence to use their data for good. They deliver the trusted and accurate data foundation needed to unlock value in data and use it to drive business outcomes.

ClearCore, their industry-leading technology, was developed alongside police analysts to find more links than ever within critical, complex, and disconnected data. They've helped provide real world impact across the public and private sector, including identifying vulnerable people, detecting fraud, and delivering multi-million-pound savings.

www.infoshare-is.com

LinkedIn Company Page

Sagacity works with some of the country's leading brands to support them in making informed decisions powered by the intelligent use of data. Sagacity believes that responsible data should be at the heart of every organisation, and helps its clients to transform their customer data into a structure they can use, enabling them to improve it, make sense of it and drive value from it.

Sagacity has delivered consistent success and ROI for its clients through new customer acquisitions and existing customer retention, customer management, onboarding, and ongoing relationship development to ensure that every consumer is treated like an individual. Data can reveal the full picture and Sagacity join the dots.

www.sagacitysolutions.co.uk

LinkedIn Company Page