Brief

5

Empowering society to overcome bias

Bias—whether intentional or unconscious—can seep into the design and delivery of services, often resulting in harm, exclusion, or inequity. From healthcare systems overlooking certain populations to financial services denying fair opportunities, biased inputs and processes can lead to discriminatory outcomes.

As society becomes increasingly dependent on technology and data-driven solutions, including AI, there is a pressing need to address the biases embedded in traditional services as well as services enabled by the latest technology. This challenge is about taking a step back, recognising the sources of bias in various industries, and designing a service that empowers service users through bias literacy. A service that gives agency to any service user, individual or group, in a private or business context, to prevent harm while fostering inclusion, fairness, and dignity for all.

What is bias

Bias refers to a systematic inclination or prejudice for or against something, someone, or a group, which often leads to unfair outcomes. Simply put, bias is a tendency to favour or disfavour someone or something unfairly, often leading to unequal or unjust outcomes.

Biases manifest in various forms, ranging from individual cognition to structural inequities, and they are deeply embedded in societal norms, historical processes, and cultural systems. The historical roots of bias are multifaceted. Psychological studies in the 1970s revealed that human decision-making often deviates from rationality due to biases like anchoring or availability heuristics. Meanwhile, sociologists have shown how systemic bias originates from historical discrimination and power dynamics. For example, historical policies and practices have created structural inequities that persist in economic and housing data used today, inadvertently feeding biases into modern AI systems.

Cognitive psychologists Amos Tversky and Daniel Kahneman have written about how bias and prejudice affect our judgement. They laid the foundation of their work in the paper ‘Judgement Under Uncertainty: Heuristics and Biases’ in the 1970s which goes into various heuristics and related biases, and how they help or hinder our decision-making capability. Also, in the book ‘Thinking, Fast and Slow’ (2011), Kahneman explores cognitive biases, identifying how mental shortcuts, or heuristics, often lead to systematic errors in judgment.

Bias can be broadly categorised into several types. Cognitive biases are errors in individual thinking processes, such as confirmation bias, where individuals favour information that aligns with their pre-existing beliefs. Implicit biases, on the other hand, are unconscious attitudes or

stereotypes that affect understanding, actions, and decisions, often perpetuating social inequalities. Sociological biases encompass systemic inequities, such as institutional racism or sexism, embedded within laws, policies, and societal structures. In technology, algorithmic bias occurs when data or design choices in artificial systems lead to skewed outcomes, often replicating human prejudices.

Recognising the forms and history of bias is critical for addressing its impacts. Whether through educational tools that raise awareness of implicit bias, policies that dismantle systemic discrimination, or technological solutions that ensure fairness in AI, tackling bias requires a multidisciplinary approach rooted in understanding its diverse manifestations.

Bias and the role of AI

Now that we have already mentioned AI, let's dive into Artificial Intelligence (AI) as a growing technology with widespread applications, including machine learning (ML). AI refers to the simulation of human intelligence by machines, while ML, a subset of AI, focuses on algorithms that allow machines to learn and adapt based on data rather than predefined instructions. This technology is used in many areas, from recommending products online to assisting in medical diagnoses and enabling autonomous vehicles. Its increasing integration into daily life is reshaping industries and influencing how decisions are made.

Because of its growing usage, AI is transforming industries, influencing decision-making, and shaping the way we interact with technology. However, AI is not immune to human flaws. The data used to train AI systems often reflects the biases present in our society today, meaning that these systems inherit and reinforce existing imbalances. When AI systems are built on biased data or flawed processes, they perpetuate and amplify societal inequalities.

These biases can manifest in subtle ways or in deeply consequential ways, undermining fairness, widening disparities, and reinforcing systemic inequities. As we move toward agentic AI systems capable of acting autonomously, the risks of bias become even more pronounced. These systems, making decisions without direct human intervention, can compound biases in unpredictable ways, intensifying the potential for unintended consequences.

For instance, voice assistants like Alexa or Siri have often performed better at recognising male voices over female ones, reflecting gender biases in their training data. In more critical areas, such as loan approvals, biased datasets have led to discriminatory outcomes, like rejecting applicants disproportionately from minority groups. These issues highlight the importance of ensuring AI systems and other technology used in services are developed and deployed in ways that prioritise fairness and inclusivity.

Understanding the problem in several industries

As mentioned earlier, bias manifests in various ways across industries, with serious implications for individuals and communities. Bias occurs when data, algorithms, or design decisions lead to unfair or skewed outcomes. Here are a few examples:

- Public Services: Digital platforms for social benefits may exclude users without internet access or language proficiency, reinforcing inequality.
- Healthcare: Medical protocols or diagnostic systems often perform worse for minority groups, perpetuating disparities in care. AI tools used to predict health outcomes often perform worse for underrepresented groups because of biased datasets. For example, some systems underestimate the severity of illnesses in women and minority patients.

- **Education**: Admission and grading systems can disadvantage students from certain socio-economic backgrounds or regions.
- **Finance**: Credit assessments disproportionately deny loans to underrepresented groups, limiting their access to economic mobility.
- **Retail**: Personalisation algorithms may alienate diverse customers by relying on narrow datasets or stereotypical preferences.
- **Recruitment**: A hiring algorithm designed to screen resumes disproportionately favoured men because it was trained on historical data where male candidates were overrepresented.

These biases undermine trust and, worse, can exacerbate inequalities. They also pose an ethical and operational challenge for businesses, governments, and individuals seeking to use data, algorithms, AI and any other technology responsibly.

The IBM Challenge: Empowering society to overcome bias

The design, development and delivery of services are often filled with biases. As a result, they may exclude people and cause harm. We ask you to design a service that empowers service users to actively engage with and navigate the biases embedded in systems and services. A solution that helps users identify biases and mitigate their impact on decision-making, experiences, or outcomes. In other words, a service that actively identifies and addresses bias, prevents harm, and promotes equity and inclusion in one of the following fields.

In the context of bias literacy, this could mean enabling them to:

- **Recognise bias**: Helping users detect where and how biases might be influencing their experiences within a service (e.g., understanding how an AI-powered loan application might evaluate them differently based on demographics).
- **Challenge bias**: Equipping users with the ability to question or confront these biases, whether by advocating for transparency, requesting fairer procedures, or raising awareness of inequities.
- **Navigate systems**: Providing agency through a service (set of tools) that allows users to make informed decisions, sidestep discriminatory processes, or demand equitable treatment (e.g., offering alternatives to biased systems or creating feedback mechanisms).

Your service won't allow individuals to completely eliminate systemic bias themselves but rather empower them to mitigate its personal and societal impacts through informed action and advocacy.

Choose your industry
You can choose one of the following industries to dive into biases. For each, we give you an example, but of course, we encourage you to find your own local problem to tackle:

1. **Retail**: Fair personalisation
Example: A service that ensures personalised marketing and shopping experiences are not exclusionary or harmful due to biased algorithms or limited datasets, such as a service that includes an audit feature allowing users to see and adjust how their preferences are inferred, promoting fairer outcomes for diverse consumer groups.

2. **Finance**: Accessible credit scoring
Example: A service that addresses bias in financial decision-making tools, such as credit scoring or loan approvals, ensuring fairness across demographics, such as a service that provides alternative credit assessment methods (e.g., rent or utility payments) to offer fairer financial access for underbanked populations.

3. **Transportation**: Accessible urban mobility
Example: A service that ensures urban transportation systems address the needs of all users, reducing bias in route planning or accessibility, such as a city-wide mobility app that uses AI to identify underserved areas and prioritises equitable service improvements.

4. **Legal**: Transparent justice
Example: A service that addresses biases in legal decisions, such as AI tools used for sentencing or bail recommendations, promoting transparency and accountability, such as a service for public defenders to analyse and challenge biased outcomes in predictive policing or legal algorithms.

5. **Healthcare**: Equitable diagnostics
Example: A service that addresses health inequities by identifying and reducing biases in medical diagnostics, patient care, or health data collection, for instance, an AI-powered platform that flags potential biases in diagnostic recommendations and provides insights to healthcare providers to improve patient outcomes for underrepresented groups.

6. **Education**: Fair admissions and assessment
Example: A service that reduces bias in educational admissions, testing, or personalised learning systems, ensuring equity for students from all backgrounds, such as a dashboard for educators to monitor and address potential biases in AI-driven grading systems, with tools for adapting to diverse learning styles.

7. **Manufacturing**: Equitable product design
Example: A service that ensures manufacturing processes consider diverse user needs and reduce potential biases, such as a platform that helps designers and engineers evaluate the inclusivity of their products and analyse prototypes or final products for accessibility, usability across demographics, or unintended cultural biases, fostering fair and inclusive outcomes in mass-produced goods.

While AI is used as an example, this challenge focuses on bias as a broader issue and encourages you to think critically about harm in both digital and non-digital service ecosystems. At the same time, AI may hold potential to be part of the solution, offering tools and capabilities to identify, mitigate, and even prevent biases when designed and applied thoughtfully.

Objectives
Your service should aim to:
- **Recognise sources of bias** in inputs (e.g., data, policies, processes) and outcomes.
- **Prevent exclusion** or harm for underserved or marginalised groups.
- **Empower users** to have greater agency in how services are delivered and experienced.
- **Promote equity** to ensure fair access and representation for all user groups.
- **Prevent harm** by actively mitigating the negative impacts of biased processes or outcomes.
- **Empower users** by enabling individuals to contribute to, shape, or challenge service outcomes.
- **Build trust** and foster confidence through transparency, inclusivity, and ethical practices in the service.

Design principles
Your design must adhere to these principles:
- **Design for inclusivity**: Incorporate diverse perspectives and needs throughout the design process to ensure accessibility for all individuals.
- **Design for transparency**: Clearly communicate how decisions are made within the service, fostering accountability and trust.
- **Design for agency**: Empower users by providing them with control over how they interact with and influence the service.
- **Design for adaptability**: Create solutions that are locally relevant while scalable to diverse contexts, ensuring long-term impact and usability.

Questions to consider
- What are the key **sources of bias** in this service area?
- How might these biases **result in exclusion** or harm for certain groups?
- What tools, frameworks, or methods can be used to **uncover and address these biases**?
- How can the service **empower users** to influence its design or outcomes?
- How can your service balance **cultural and societal diversity** while addressing global issues?
- How can you **foster trust and engagement** with your service among diverse user groups?

IBM's AI Ethics Principles

Explainability
Good design does not sacrifice transparency in creating a seamless experience.

Fairness
Properly calibrated, AI can assist humans in making choices more fairly.

Robustness
As systems are employed to make crucial decisions, AI must be secure and robust.

Transparency
Transparency reinforces trust, and the best way to promote transparency is through disclosure.

Privacy
AI systems must prioritize and safeguard consumers' privacy and data rights.

Supporting resources
To guide your solution, consider these resources:

- Articles on bias in service delivery across industries.
- Frameworks for inclusive design and equity-centred practices.
- Examples of services that have successfully mitigated bias or prevented harm.

Here are a few to get you started:

- How heuristics impact our judgment ⌕ - A short animation that explains the ideas presented by Kahneman and Tversky in their 1974 paper looking at three heuristics commonly employed: representativeness, availability and anchoring.
- Fairness: Types of bias ⌕ - Inventory that provides just a small selection of biases that are often uncovered in machine learning datasets

Additional IBM tools and frameworks to explore and guide your solution:

- AI Fairness 360 Toolkit ⌕: Tools to detect and mitigate bias in machine learning models.
- IBM AI Ethics Framework ⌕: Everyday ethics for designing and deploying AI.
- Research on AI ⌕: Learn about IBM's AI advancements and the Cambridge lab.
- IBM Models and Platforms ⌕: Leverage platforms like IBM Watson and Granite.

Your mission
This is your opportunity to address one of the most pressing challenges in service design: bias. We cannot change the services that are already there, sometimes for many years, and often based on unintended biases. However, you can **empower service users to identify, challenge, and overcome these biases**. By recognising and dealing with biases in services, you have the potential to help service users and therefore create equitable and empowering experiences for all. The solution you design today can set the standard for fairness and inclusivity in the future. What will you design to ensure that no one is left behind?

About IBM and IBM Research

This is a challenge offered to you by IBM. IBM is a global technology leader with over a century of innovation history. As a pioneer in artificial intelligence, IBM is committed to developing technology that is inclusive, transparent, and aligned with societal values. IBM drives meaningful innovation through partnerships with industries, governments, and communities worldwide, creating solutions that improve lives and businesses.

IBM Research, the company's advanced research division, focuses on pushing the boundaries of technology. With a lab in Cambridge, IBM Research is at the forefront of AI ethics, studying bias, fairness, and the societal impacts of AI. Their work informs the development of IBM's tools, such as AI Fairness 360, and ensures that ethical considerations remain central to AI innovation.

Together, IBM and IBM Research exemplify a shared vision: advancing technology that works for all.