1. ﻿**How do we address bias in data science models? Example: What steps can be taken to mitigate biases in healthcare AI?**

Bias in AI can significantly impact individuals by influencing decisions like school admissions, job hirings, and rental applications. These systems often exhibit biases due to their programming and training data, which may lack diversity and underrepresent certain genders or ethnic groups. This issue is especially critical in healthcare AI, where biased models can result in misdiagnoses and unequal treatment of patients. To mitigate this issue, it is essential to use balanced datasets and exclude sensitive attributes, such as gender or race, during the training phase. Implementing fairness metrics to evaluate model outputs is also vital. Techniques like counterfactual fairness allow us to test predictions for bias by altering sensitive variables. Furthermore, model validation in different contexts is also important for ensuring fairness and accuracy.

1. ﻿**What are the limits of informed consent in big data applications? Example: How do we ensure consent is meaningful in large-scale social media data use?**

By ‘informed consent’ we have in mind the sense articulated by Tom Beauchamp, who says—in the context of medicine, clinical practices, and biomedical research—that ‘A person gives an informed consent…if and only if the person, with substantial understanding and in substantial absence of control by others, intentionally authorizes a health professional to do something’ (2011, p. 517–518). In the context of big data, informed consent can become limited, as users often lack a clear understanding of how their data will be utilized, particularly within complex systems such as social media. To ensure that consent is meaningful, organizations should try to simplify their data usage policies, incorporating summaries or highlighted text for more clarity, and providing ongoing opportunities for users to withdraw their consent. It is also essential to maintain transparency regarding data-sharing practices and to seek explicit, informed approval for each new purpose to advocate ethical standards.

1. ﻿**How can transparency and interpretability be balanced with complexity? Example: Should there be mandatory explainability for AI systems impacting financial decisions?**

Modern AI models, like deep learning, can be very complex and hard to understand. When these systems are used for important decisions, such as job hirings or loan approvals, it is important to be transparent to ensure accountability. The transparency problem occurs when companies are reluctant to share internal processes or when complex algorithms prevent understanding of the process. Users have a right to explanations about data use but achieving informed consent can be challenging. While excessive transparency may overwhelm users; context and potential harm should guide the appropriate level. An example highlights this issue: a teacher received a poor performance score from an algorithm without knowing why cannot really benefit the teacher, students or the education system. This is underscoring the importance of transparency in understanding decision-making processes. Techniques for model explainability, including simpler models or local explanations like XAI, LIME and SHAP, help make complex systems easier to understand. In financial AI, mandatory explainability could require these systems to give users clear reasons for their decisions, making sure those decisions are fair and justifiable.

1. ﻿**How should organizations disclose the use of AI tools like ChatGPT?**

It's necessary to disclose the use of AI tools like ChatGPT for ethical transparency. Organizations must communicate when and how they use AI in areas such as customer service, decision-making, or content creation. Informing users about AI-generated content helps them recognize its limitations, including the absence of human judgment. Ethical guidelines recommend human oversight in situations where AI's limitations could negatively impact users.

﻿**Case Study 1: Biased Algorithms**

In 2014, Amazon developed an experimental AI hiring tool aimed at streamlining the recruitment process by evaluating and ranking job applicants. However, by 2015, it became apparent that the system displayed a significant gender bias, systematically disadvantaging female candidates. This bias stemmed from the training data used: the AI was trained on résumés submitted to Amazon over the course of a decade, during which the tech industry—and Amazon's own workforce—was predominantly male. As a result, the AI learned to favor male candidates, penalizing résumés that contained terms like "women's," such as "women's chess club captain," and downgrading graduates from all-women's colleges. Despite efforts to adjust the algorithm for greater gender neutrality, these biases persisted, ultimately leading Amazon to discontinue the tool by 2017. We cannot assume that artificial intelligence is inherently unbiased. When these systems are trained on biased data, the algorithms also inherit that bias. If unfair AI hiring programs are not identified and addressed before implementation, they will continue to reinforce existing diversity issues in the workplace rather than resolve them.

﻿**Case Study 2: Data Privacy Breach**

The Cambridge Analytica scandal started in 2018 when it was discovered that the political consulting firm had illegally taken personal data from millions of Facebook users without their permission. They gathered this data through an app called "This Is Your Digital Life," created by Aleksandr Kogan. This app collected information not just from the users who took the survey but also from their Facebook friends, affecting up to 87 million people. The event showed serious problems with data privacy and how consent was handled. Many users did not know their information was used for political profiling and targeted ads. Due to public backlash and legal pressure, Facebook faced major fines and lawsuits. In December 2024, Meta Platforms, Facebook's parent company, agreed to pay $50 million to settle a case with Australia's privacy watchdog over sharing user data improperly. This case highlighted the need for strong regulations on data protection, leading to the creation of the General Data Protection Regulation (GDPR) in the European Union in May 2018. The GDPR set clear rules for how businesses and organizations can collect and use personal data. It requires that users give clear consent for their data to be used and gives people more control over their personal information.

﻿**Case Study 3: Facial Recognition Technology**

Facial recognition technology (FRT) is used in Australian venues, raising serious privacy concerns. About 49% of Australians are unaware that venues like Melbourne's AAMI Park and Sydney's Allianz Stadium use it. Many learned about it through news reports. Legal experts argue that current privacy laws do not effectively regulate FRT and emphasize the need for clear consent from visitors. Despite these concerns, only 7% intend to avoid venues using FRT, highlighting the urgent need for updated privacy regulations. The Australian Privacy Commissioner found that Bunnings, a major retail chain, violated privacy laws by using FRT without properly informing customers. Between 2018 and 2021, Bunnings collected facial images from hundreds of thousands of shoppers without their consent. The Commissioner deemed this practice intrusive and ordered Bunnings to stop, showing the need to balance security and privacy rights. These cases reflect a broader discussion about the ethical use of FRT, including consent and data security. As FRT becomes more common, organizations must adopt clear practices, and regulators should create guidelines to protect privacy rights.

﻿**Case Study 4: Redlining**

Redlining is a practice where banks and lenders deny loans and financial services to certain neighborhoods, often based on race or ethnicity. This term comes from the red lines drawn on maps in the 1930s, marking areas considered too risky for investment, mostly those with African American and minority residents. The Federal Housing Administration (FHA) was created in 1934 and used these maps to decide on mortgage insurance. This meant minority communities often could not get home loans. This practice led to segregation and made it harder for these communities to build wealth, as owning a home is a major way to gain financial stability in the U.S. Even after the Fair Housing Act of 1968 aimed to stop housing discrimination, the effects of redlining still exist. Neighborhoods that were once redlined continue to have lower homeownership rates and less access to loans. Studies show these areas still mostly contain minority populations and face ongoing economic issues. In the field of artificial intelligence, there is concern that these old biases can continue through algorithmic decision-making. AI systems that learn from historical data may repeat these biases, leading to unfair treatment in lending and housing. To address this, researchers are creating ways to reduce bias in AI in three stages: changing the training data before it is used, adjusting AI learning methods, and modifying AI outputs to ensure fairness. Understanding the history and effects of redlining is important for building AI systems that support fairness and do not repeat past mistakes.