In qualitative analysis, a codebook serves as a data guide, including codes, definitions, and examples from participants' responses [1]. Our codebook, reported in Table 1, supports first-cycle coding analysis, ensuring consistent coding and data interpretation among the research team.

Table 1: Codes for the first-cycle coding, along with their definitions and representative participant quotes.

Code	**Definition**	**Representative Quote**
Humans can see beyond the number	The belief that AI systems rely strictly on numerical inputs or quantifiable data and therefore cannot interpret context, nuance, or situational factors that fall outside the dataset.	"using employees to assess the risk of developing heart disease would be beneficial because a human will be better able to look at the broader picture of a person and not be limited to just the data points they are told to use."
Ability: see gray area	Humans can interpret ambiguous situations, weigh conflicting information, and make judgments in cases that are not clearly defined or rule-governed.	"In gray areas where they would use their own judgement in whether I would be at risk of non-payment like educational background, family background etc."
Flexibility: not strict with the rules	Human decision-makers can deviate from rigid procedures when needed, applying discretion to adjust decisions when strict rule-following would produce an unfair or unreasonable outcome.	"An employee who is a real person may be more likely to understand the nuances of how much a person can afford for a loan and how likely they are to miss payments, rather than strictly using an algorithm."
Ability: all factors	Humans are perceived as capable of considering a broad range of relevant information.	"Employees can use multiple sources and past experience to determine risk. Making these assessments regularly would improve their ability to make accurate judgements about risk."
Ability: individual circumstances	Humans can take into account applicants' unique life situations, contextual details, or personal histories that influence the fairness of a decision.	"Employees can use multiple sources and past experience to determine risk. Making these assessments regularly would improve their ability to make accurate judgments about risk."
Ability: not to follow the social pattern	Humans can resist or challenge stereotypes, social biases, or generalizations, instead evaluating individuals on their own circumstances.	"When the person is a minority"
Ability: adjust accordingly (not rigid)	Human decision-making is adaptable, allowing evaluators to revise their judgment based on new information, conversation, or situational understanding.	"Human employees can understand certain life situations and make acceptations."
Ability: human quality	General perception that humans possess inherently beneficial qualities (e.g., empathy, morality, common sense) that improve the fairness of decisions.	"I'm just not sure how well AI would be at accessing human qualities that would be important for the job, besides skills and qualifications."
Ability: communication skills	Humans can communicate with the candidate and oberser the communication skill.	"Employees can get a read and a feel of someone's personality, appearance, and other qualities like good communication skills"
Ability: soft skills	Humans can judge the soft skills.	"In a scenario of team dynamics, employee can assess soft skills of the applicants, which would be beneficial for assessing the quality."

Ability: interpersonal skills	Humans can judge another person's nonverbal cues—skills viewed as beyond the capabilities of AI.	"Using employees are beneficial for assessing the quality of the applicant because they are able to view the applicant personally to see if they are a great fit for the company overall."
Non-traditional financial background	Human has the ability to evaluate unique, non-traditional financial profiles.	"Using employees to assess risk is beneficial when a client has a non-traditional financial background, such as irregular income from self-employment, that standard credit models might misinterpret. Employees can consider the full financial picture and apply context, leading to a fairer and more accurate assessment."
Applicant is self-employed/ irregular income	Participants mentioned this type of applicant in their responses to describe how this group of people will benefit from human decision-makers.	"Using employees to assess risk is beneficial when a client has a non-traditional financial background, such as irregular income from self-employment, that standard credit models might misinterpret. Employees can consider the full financial picture and apply context, leading to a fairer and more accurate assessment."
Recent change: divorce	Participants mentioned several life-changing aspects that can affect their financial profile, such as divorce.	"A trained employee can use context, judgment, and humanity to make a more equitable decision when an applicant has a unique monetary circumstance that might not be fully recorded in conventional data, such as a recent separation and divorce, self-employment, or non-traditional income."
Recent change: medical issue	Participants mentioned several life-changing aspects that can affect their financial profile, such as unexpected medical cost.	"Using employees may be beneficial when assessing clients with complex or non-traditional financial histories, like small business owners, freelancers, or individuals with recent life changes (e.g., divorce, medical issues). Human employees can interpret context, ask follow-up questions, and exercise discretion where standardized algorithms might be rigid or fail to understand nuance."
Recent change: immigrant	Participants mentioned several life-changing aspects that can affect their financial profile, such as immigrant.	"Using employees is useful when evaluating applicants with non-traditional financial backgrounds ,AI such as gig workers, freelancers or immigrants ,AI who may not have standard documentation but still pose a low risk of default."
Recent change: losing a job	Participants mentioned several life-changing aspects that can affect their financial profile, such as loosing a job.	"If a client has an irregular income or recent life changes (e.g., recovering from a job loss), an employee can understand the nuance and override strict data-based decisions."
Can judge appearances	Human can utilize their intuition by looking at applicant's appearances	"The employees can better assess personality traits or appearances when it comes to making payments"

Compassion	Human can show compassion in their judgment	"A person has compassion and empathy. "
Ability: circumstance for committing the crime	Human can consider the motive of a crime.	"they can become aware of certain extenuating circumstances in the offense"
Ability: check beyond what is available	Human has the ability to gather more information beyond what is presented.	"Employees can perform personalized checks (e.g., taking to social workers, families, victims, etc) to make a more informed decision"
Ability: human behavior	A human has the ability to judge another person's behavioral pattern.	"Any situation assessing a human's likely behavioral patterns is best assessed by another human with relevant expertise."
Ability: experience	Human can utilize their experience from doing the work into their decision-making	"These are people with, one would assume, experience in doing this exact work. They have absorbed years of knowledge about who is likely to reoffend based on a number of factors, such as age, race, crime, education, etc."
Accountability	Participants noted that we can hold humans accountable for their decision-making.	"There would be some accountability in having that."
Consider: lifestyle factors	Humans have the ability to consider lifestyle factors when making decisions.	"Using employees can be beneficial in complex cases where medical history or lifestyle factors require a personalized review,AI for example, someone with a rare genetic condition or unusual risk factors that aren't well-represented in existing data."
Border picture or overall health of a person	Humans can consider various lifestyle factors that contribute to health when making decisions.	"Employees can observe an overall picture of someone's health."
Complicated/ complex health profile	Humans has the ability to evaluate complicated and complex health profile.	"Using employees would be beneficial when an applicant has a complex or rare medical history not well captured by standard metrics. For instance, someone with unusual genetic markers or a rare condition could be better assessed by a medical expert who can interpret specialist notes and cross-disciplinary insights."
Unique health condition	Humans has the ability to evaluate unique health profile.	"If there are unusual health conditions or outlying circumstances, it could be very helpful to have a real human employee's expertise and opinion."
Rare health condition	Humans has the ability to evaluate rare health conditions.	"Using employees would be beneficial when an applicant has a complex or rare medical history not well captured by standard metrics. For instance, someone with unusual genetic markers or a rare condition could be better assessed by a medical expert who can interpret specialist notes and cross-disciplinary insights."
Consider family history	Humans have the ability to consider family medical history when evaluating the risk of certain diseases.	"Using employees is beneficial when personalised interviews or nuanced judgement is needed to assess lifestyle factors and family history that automated systems might overlook."

Customer service	Humans have the ability to judge who will perform better in a customer service roll.	"Customer service is a good example....the way an employee interacts with a customer and their knowledge."
Specialized skills	A specialized individual with specific skills has the ability to assess those specialized skills.	"Employee-led assessments are most valuable when evaluating specialized technical skills, e.g., debugging legacy code that only experienced team members can properly test. They excel at identifying culture-add candidates who bring fresh but compatible perspectives, beyond what resumes show. This approach works best when employees use structured rubrics to reduce bias while leveraging their hands-on role knowledge."
Candidates for leadership	Human can evaluate the leadership skill	"When assessing candidates for leadership or communication-heavy roles, employees can better evaluate soft skills during interviews or assess cultural fit something AI may not detect from a resume alone."
Candidate to work on a new invention	Human can evaluate the skills that need to work on an innovative project	"when there's a new invention and others are needed to try it out because their opinions matter to assess quality."
Empathy can influence the decision	Human empathy can influence the decision making either being sympathetic or being harsh.	"If employees are considerate and will be lenient to cover more people"
Human empathy and reasoning	Human has the empathy and reasoning and can utilize then in decision making	"If employees are considerate and will be lenient to cover more people"
Ability: evaluate team capability	Humans can better assess a team's capabilities and dynamics than an AI system.	"In a scenario of team dynamics, employee can assess soft skills of the applicants, which would be beneficial for assessing the quality."
Ability: applicants fit within the company culture	This code captures the view that humans are better equipped to judge whether applicants fit the company's culture and values.	"To determine how well an applicant would fit within the culture."
Community based benefits	Human decision maker can make decision that benefits to the broader company, not just individual outcomes.	"Employees understand the needs of the company and the priorities set by management. They can also use their own discretion when doing the assessment, meaning that I can potentially grab their attention in unexpected ways."
Know the company's need	Humans understand the company's specific needs and priorities better than an AI system.	"Employees understand the needs of the company and the priorities set by management. They can also use their own discretion when doing the assessment, meaning that I can potentially grab their attention in unexpected ways."
Interview or getting explanation or Explainability	Participants value the interactive nature of human decision-making, where applicants can ask questions, clarify misunderstandings, or receive explanations directly from the decision-maker.	"Employees can talk with the person and get a better understanding of what their needs are."

Flexibility and nuance	Humans are perceived as capable of interpreting subtle cues, balancing conflicting considerations, and applying nuanced judgment in complex cases.	"Employees may have a deeper understanding of individual prisoners' backgrounds and motivations. For instance, if a prisoner has demonstrated significant behavioral change or rehabilitation efforts, an employee might recognize this and adjust their assessment accordingly."
Human intuition	Humans can draw on instinct, experience, and tacit knowledge that cannot be easily formalized or replicated by algorithmic systems. bar	"Use of employees would be beneficial when risk assessment requires human intuition or empathy, such as evaluating complex financial situations that might not necessarily be addressed through data alone."
Personal experience with the prisoners	Humans draw on personal experience with prisoners, giving them contextual insight that AI lacks.	"Employees are around the prisoners for a long time so it is easy for them to make the decisions."
Real-life experience (doing the job) to judge others	Humans use real-life, on-the-job experience to more accurately judge others' suitability.	"Employees can use multiple sources and past experience to determine risk. Making these assessments regularly would improve their ability to make accurate judgements about risk."
Experience or expertise in subject matter	Humans rely on subject-matter experience or expertise to make more informed and contextually grounded judgments.	"Employees can use multiple sources and past experience to determine risk. Making these assessments regularly would improve their ability to make accurate judgements about risk."
Experience as a human	Humans draw on lived experience to make judgments in ways AI cannot.	"Because they have human experiences to rely upon "
This is the norm or way to do	Participants perceived that human will do the specific task as this is the norm of doing the task.	"This is the standard way of hiring employees for ages. Whether it be by interview or resume, employees assessing the quality of candidates is how you hire people."
Trust in humans	Participants showed a general trust on human decision maker.	"Because they have human experiences to rely upon "
Non-bias or less-bias	This code reflects participants' belief that humans can exercise less bias or correct for bias in ways AI cannot.	"I would think that using employees is likely to be less bias and more open."
Assess a human by a fellow human	Human judgment is better suited for evaluating another person's character, behaviour, or potential than an AI system.	"It makes sense for human applicants to be assessed by fellow humans. They would know more about what would make an applicant successful in the role."
If: does not discriminate	Participants trust human decision-makers only when they act fairly and avoid discrimination.	"If the employee does not discriminate in any way. A person can determine the quality."
If: has all the info needed	Participants trust human decision-makers only when they has all the information necessary to make such a decision.	"If the employees are well trained and use data to back up their decisions."
If: has proper training	Participants trust human decision-makers only when they have the proper training to perform the task.	"If the employees are well trained and use data to back up their decisions."
Straightforward cases	Human could be beneficial for very straightforwards criminal cases.	"Could be beneficial when you are reviewing petty crimes and the length is not that impactful to the criminal"

Bias and discrimination	Humans may introduce bias or discrimination that results in unfair decisions.	“This might be harmful as a result of bias”
Bias: ethnicity	Humans may show bias based on ethnicity, leading to unequal treatment.	“Using employees could be harmful if personal bias or stereotyping creeps in. For example, an employee may subconsciously judge a person’s risk based on weight or ethnicity, even if those factors are not the most predictive. This could lead to unfair premium rates.”
Bias: gender	Humans may show bias based on gender, leading to unequal treatment.	“They could have unconscious implicit biases with regards to race, sexuality, religion, etc., that could affect their judgment."
Bias: sex	Humans may show bias based on sex, leading to unequal treatment.	“They could have unconscious implicit biases with regards to race, sexuality, religion, etc., that could affect their judgment.”
Bias: race	Humans may show bias based on race, leading to unequal treatment.	“They could have unconscious implicit biases with regards to race, sexuality, religion, etc., that could affect their judgment.”
Bias: community or neighborhood	Humans may show bias based on community, zip code, neighborhood, leading to unequal treatment.	Let’s say a staff member has unconscious bias against people from a certain background like race or neighborhood. That could cause them to wrongly label someone as "high risk" even if the person is actually low risk, leading to a longer and unfair sentence.
Bias: religion	Humans may show bias based on religion, leading to unequal treatment.	“They could have unconscious implicit biases with regards to race, sexuality, religion, etc., that could affect their judgment.”
Unconscious bias: stereotype	Humans may show unconscious bias based on stereotype, leading to unequal treatment.	“A harmful scenario would be when an employee’s personal biases, conscious or unconscious affect their judgment. For instance, if an employee stereotypes applicants based on their ethnicity or socioeconomic background, it could lead to unfairly high interest rates or loan denial for capable clients.”
Bias: know the offender	Humans may show bias when they personally know the offender, affecting impartiality.	“Bribery comes to mind. If an inmate wants out and he has the money, bribery can always work. Also, if they get to know the employee that will make the decision, they can treat them better than they normally would treat someone to get on their good side.”
Bias: know the victim	Humans may show bias towards the offender when they personally know the victim, compromising fairness.	“If the person committed a white collar crime against the same community and the panel knew victims or felt close to them”
Bias: based on crime type	Humans may show bias based on crime type.	“employees may be biased towards certain crimes”
Bias: marital status	Humans may show bias based on marital status, leading to unequal treatment.	“If the employee is prejudice towards certain applicants such as race, sex, marital status”

Bias: socioeconomic	Humans may show bias based on socioeconomic background, leading to unequal treatment.	"When dealing with high profile clients"
Bias: know the applicant	Humans may show bias if they personally know the applicant, leading to unequal treatment.	"But anyone else would likely lean toward friends and possibly sob stories and cost the bank money."
judge based on appearance	Humans may show bias based on appearance, leading to unequal treatment.	"An employees personal bias causes them to have high rates for a low risk client based on irrelevant factors. Maybe like through appearances."
Bias: fear of replacement by the applicant	Humans may show bias if they fear being replaced by the applicant, influencing their judgment.	"There can be an inherent bias or fear of being redundant in their position, this could lead them to vouch for a less skilled hire to assure job security."
Favoritism: famous schools	Humans may show bias by favouring applicants from prestigious or well-known schools.	"Employees might have hidden biases, like liking applicants from famous schools or similar backgrounds. Because of this, they might miss out on great candidates who took different paths or come from less-represented groups."
Favoritism: certain ethnicity	Humans may show favoritism toward certain ethnic groups, leading to unequal treatment.	"When the interview and interviewee are members of different racial, ethnic or socioeconomic groups; or have different work backgrounds/histories."
Favoritism: towards similar background	Humans may show favoritism toward applicant with similar background, leading to unequal treatment.	"Harmful if employees show bias or favor people like themselves."
Emotional interference make unfair sentencing	Humans may let emotions influence their decisions, reducing objectivity.	"If the employees are biased or have an emotional grudge or even unintentional bias they may judge more harshly or non objectively and say yes or no when not warranted"
Empathy can interference sentencing	Humans empathy can influence their decisions, making lower interest rate, or lighter sentencing.	"A person has compassion and empathy."
Human intuition influence the decision	Human intuition may override evidence or evaluating criteria, leading to inconsistent decisions.	"if they are doing it out of their one intuition they may be biased, and may give inconsistent assessments."
Recruiter having a bad day	A recruiter's mood or having a bad day may negatively influence their judgment.	"if an employee is simply having a bad day, they may pass a good candidate over just based on their negative mood"
Inconsistency	Human decisions can be inconsistent across similar cases or situations.	"Employees may introduce bias or inconsistency in their judgments, especially if they are influenced by unconscious stereotypes or lack standardized assessment guidelines. This could lead to unfair outcomes, especially for clients from minority or disadvantaged backgrounds."
Error prone	Human are prone to make mistakes.	"An employee might have bias or make a mistake on the paperwork."

Human: tiredness	Tiredness can impair human judgment and reduce decision quality.	"When an employee overlook important information due to being tired or not paying attention."
Not rule based	Humans may deviate from rules or standards, leading to unpredictable decisions.	"...This lack of standardization and potential for human error could result in some low-risk individuals being unfairly labeled as high-risk and receiving longer sentences."
Lack of expertise or experience	Lack of experience or expertise can lead to poor decision-making.	"A harmful situation would be if untrained employees were responsible for assessing heart disease risk without proper medical knowledge leading to inaccurate evaluations."
Employee might need to meet quota	Humans may make decision in a way to meet quotas or targets rather than evaluate cases fairly.	"Unnecessary premium added to the quote"
Corruption	Humans may engage in corruption, compromising fairness and integrity.	"The only thing I can think of is that the employee makes a human mistake when assessing the risk, or if they bribe the employee."
Conflict of interest	Humans may face conflicts of interest that undermine impartial decision-making.	"If the new candidate would be in direct competition for the other employee's jobs, then the original workers should not be consulted to judge the new person"
No trust in human	Participants expressed a general distrust on human, without specific reasoning.	All situations.
Cut down the cost of human labour	AI can reduce costs by cutting down the need for human labour in decision-making processes.	"We could cut cost on labor if we implemented AI"
None of the options is fair	Mentioning that all of the option has some sort of unfairness.	"None of these options are fair. If there is an issue that is known, then you cannot implement that technology whilst intentionally impacting the actual lives of others. It goes against our core democratic rights."
Negative attitudes towards AI	Participants showed a negative attitude in terms of using the AI model, some also mentioned not using AI for the task	"I don't think there's really a good answer here - if there is bias in the program, it should not be used! Someone will be treated unfairly regardless."
Inaccuracy leads to a negative social outcome	Due to inaccurate prison sentencing, high-risk prisoners will get out of prison, which is not good for society.	"Bad people could end up doing bad things to more people."
One group is getting unfair benefits	Participants mention that brunettes are getting an unfair advantage in the given task.	"Brunettes get unfairly shorter sentences, harming fairness for others."
Negative effect on the company or institution	The company might get sued, be subject to public backlash or people losing trust in them	"They could get sued"

Regulatory mention	Participants mentioned regulations such as HIPAA or only mentioned the company might get sued due to treating people unfairly or giving inaccurate predictions	"They could get sued"
Negative impact on low-risk or highly qualified group	Due to inaccuracies, the low-risk group in prison, mortgage, and insurance scenarios, and highly qualified candidates in job scenarios will receive high sentences, high interest rates, or high premiums, or be marked as low-potential candidates.	"Prisoners with minor crimes receiving unfair sentences. And brunettes walking free from severe felony charges."
Do not input such a feature	Participants mentioned not to input irrelevant features into the AI	"I don't think hair color should be included in data set at all."
Negative impacts on the client	Participants mentioned how	"Incorrect sentences are unfair and may be a violation of someone's rights. This could lead to prisoners having their sentences thrown out and/or possible civil penalties."
Inaccuracy	Mentioning the inaccuracy of the given task	"incorrect interest rates"
Negative impact on the innocent group	In the prison scenario, due to inaccuracy innocent people will suffer.	"This option would unnecessarily make an innocent person spend longer time in prison."
Balancing fairness and accuracy	Without choosing one metric over the other, participants emphasized balancing the two.	"This balances fairness and accuracy. Completely removing bias at the cost of highly incorrect sentencing would be unjust."
Overall better outcome	Participants suggested that their suggestion provides an overall better outcome.	"The AI model does not favor anyone and still provides accurate suggestions"
Acceptance of the tradeoff	This code can be used when participants are willing to sacrifice one metric for achieving another, and when participants show acceptance of a balance between the two metrics.	"If the AI doesn't favor brunettes, then more people are negatively affected by incorrect rates."
A decision based on appearance is unfair	Any decision or prediction based on appearance is perceived as unfair to the participants	"Should TechForLegal not deal with its bias but hide it instead, prisoners may face unfair sentence decisions based only on their appearance, which could result in injustice."
Catch 22	Nothing is good	"It prevents fairness. It either favors one side, or provides info that is too incorrect."
No good options	Participants mentioned that none of the options are good to choose from one	"I don't think there's really a good answer here - if there is bias in the program, it should not be used! Someone will be treated unfairly regardless."
Human assessment after AI: brunettes	Expressing having human assessment after the AI prediction	"Every sentence for brunettes must be human reviewed, taking extra time and money."
Prioritize correct assessment	Expressing the importance of receiving accurate prediction	"anything othervthan my pick will be a popssibility of an incorrect prison sentence. and we need the correct prison sentence."
Prioritize fairness	Expressing the importance of receiving fair prediction	"Any other option I believe has too much bias in regards to brunettes."

Introduce a new bias	AI might introduce a new kind of bias	"it could also start creating a bias against individuals by associating them based on characteristics and not personality/ record"
Favoring one group is better than marginalizing the group	Mentioning that showing favoritism is better than having negative bias towards any group	"It would provide incorect prison sentences. I think favoring one group is better than marginilising the group. I would rather it be lenient than harsh. If it goes with any of the other options it would end up being harsher than it should have beein otherwise."
Inaccurate results, accept bias.	Want accurate prediction therefore, by default accepting bias.	"I want an accurate prison sentence if it is lighter on brunettes fine but I don't want incorrect sentences at all."
No bias, prefer inaccuracy for everyone	Want fairness in prediction therefore, by having inaccuracy for everyone.	"If I were an applicant, I would suggest that MedForYou prioritize fairness and remove the bias toward brunettes, even if it means some low-risk applicants may temporarily receive slightly higher premiums."
Monetary loss for the client	In insurance and mortgage scenarios, if low-risk participants are misclassified as high risk, they might pay more than they should.	"Low risk applicants pay higher interest rates."
Monetary loss for the company	In insurance and mortgage scenario, if the company give low interest rate or premium to high risk applicant, the company will loose money,	"Monetary loss for the company due to inaccuracy - giving lower rate to the higher risk"
losing trust: inaccuracy	Due to receiving inaccurate result, people will lose trust in the company	"If InsureYou constantly provides incorrect interest rates. it would lose customers as they would find better deals elsewhere."
Inaccuracy: harm the quality of hire	In job scenario, an inaccurate prediction about who is highly potential, it will harm the overall quality of the hire	"Ignoring this approach could discourage talented candidates from applying, reducing overall workforce quality"
Losing trust: bias	Due to receiving unfair result, people will lose trust in the company	"If RecruitTalent decides to preserve accuracy at the expense of fairness, the system would continue to recommend low potential brunettes over high-potential non-brunettes. This not only results in poor hiring decisions but could also erode trust in the company, attract public criticism, and expose TechForYou to reputational or legal risk due to discriminatory practices."
Negative impact on the diversity of hire	Due to favouring one group, in the hiring scenario, it will effect the diversity of the hire.	"If AI is prompted to choose extremes, then it allows less chance of diverse candidates of being chose."
Fairness is more important than Membership Inference (MI)	Participants express that ensuring fairness is more important than protecting MI in this context	"If the AI does not favor brunettes, then the system will be better, as the revealing the wealth status of the inmates may not have a major negative impact on them."

Revealing information leads to no harm / Membership information is irrelevant here	Revealing this specific information is perceived as not harmful and irrelevant in the decision-making context	"I don't see as how revealing if the applicant is wealthy is relevant. Their finanacial status shouldn't have any bearing on their sentence or chance to reoffend."
Unwanted information reveal might effect the outcome	Participant perceived that MI might effect the decision-making	"If the AI strongly favours brunettes and reveals a prisoner's wealth status, it could create an unfair advantage for wealthier prisoners, who may receive preferential treatment regardless of their actual risk of re-offending."
Prioritize membership privacy	Protecting MI is prioritized by the responder	"If the AI strongly favours brunettes and reveals a prisoner's wealth status, it could create an unfair advantage for wealthier prisoners, who may receive preferential treatment regardless of their actual risk of re-offending."
Privacy violation	Revealing MI is a violation of privacy	"It could be a data privacy concern and create a bias against wealthy"
Biased based on MI	Decision might be biased based on the membership status	"Someone could be given the wrong prison sentence as well as be labeled as someone with a lot of money which could create bias"
Less bias, accept privacy risk	Participants want fairness in the decision-making, therefore by default accepting the privacy risk	"I'm not sure what the harm is in just revealing that an applicant was wealthy, in fact it maybe a good think as you could then see if they received special treatment."
Prioritize Fairness	Ensuring fairness in decision making is prioritized by the responder	"Again, the issue is bias for brunettes and varying degrees of revealing if a prisoner is wealthy or not which I fail to see how that matters in anything. It matters more to not have any bias. The fact they aren't binning the whole thing is the real issue. It's not like there is only 1 AI option."
Accuracy should be focused	Participants expressed that the correct assessment should be focused in decision making	"There is opportunity for incorrect sentencing. I think privacy of wealth is not enough risk compared to incorrect sentencing or bias."
Biased result is inaccurate	A favourable outcome is also an inaccurate outcome as it does not reflect the true prediction	"There is opportunity for incorrect sentencing. I think privacy of wealth is not enough risk compared to incorrect sentencing or bias."
Balancing membership privacy and fairness	Participants chose the option that balances the membership privacy and fairness	"By only having a slight chance of revealing whether applicants have depression that largely (but not completely) mitigates pricing on the basis of depression rather than risk of heart disease. Strongly favoring brunettes however may result in mispricing of policies to a larger degree, resulting in a greater risk of losses for the insurance company and future price hikes."
Transparency	Participants' perceived distribution statistic revealing as a transparency metric.	"if i understand this correctly, and i'm not sure i do, it sounds like it would be a good thing. transparency when creating the data is good"

Distribution information not important	Participants' perceived distribution statistic revealing as not a priority.	"Again, we're going back to fairness. This is about crimes and the chance to reoffend not gender. Who cares if it knows how many women were in the data set. It doesn't even need a data set it needs a rule based system to operate off of."
Distribution information is good to know	Participants' perceived distribution statistic revealing as a positive thing to know about the AI.	"Once again, the best scenario is for no one to be favored at all. The other three scenarios would present biases in favoring brunettes, which is definitively unfair to all non-brunette prisoners. Furthermore, if the number of women in the data is revealed, this could be a good opportunity to make sure the data is not skewed."
Distribution information reveal might effect the outcome	Revealing distribution information can influence the decision outcome in unintended or unfair ways.	"If the AI strongly favours brunettes and reveals the number of women, this could lead to gender discrimination, where male prisoners might receive more lenient sentences, or female prisoners could be unfairly targeted."
Privacy violation	Revealing distribution information is a privacy violation.	"It could be a data privacy concern and create a bias against women"
Protect privacy, accept bias	Participants are leaned towards protecting privacy by accepting bias.	"It could be a data privacy concern and create a bias against women"
Less bias, accept distribution reveal	Participants are leaned towards unfair outcome by accepting privacy risk.	"The AI must not favor any group of individuals. There is nor risk associated with revealing the number of women in the data."
Prioritize Fairness	Participants prioritize fairness in decision-making.	"The AI must not favor any group of individuals. There is nor risk associated with revealing the number of women in the data."
Negative effect on societal institution due to bias	Bias in AI decisions can negatively impact societal institutions and reinforce structural inequalities.	"If TechForLegal's priority is protecting women rather than tackling bias in the system, then unfair sentencing based on looks might carry on, which could negatively impact the trust in the justice system."
Loosing trust due to privacy violation	Privacy violations can lead people to lose trust in the AI system and the institutions using it.	"If TechForLegal's priority is protecting women rather than tackling bias in the system, then unfair sentencing based on looks might carry on, which could negatively impact the trust in the justice system."
Accuracy should be focused	Accuracy should be prioritized to ensure decisions are reliable and minimise errors.	"Again, inaccurate sentences for brunettes. Population of women would be fine compared to this."
Balancing distribution privacy and fairness	Participants chose an option that balance both distribution privacy and fairness.	"Any other option gives too much of a chance to reveal the amount of women, or once again has too much bias towards brunettes."
Negative attitudes towards AI	Participants hold negative attitudes toward AI, affecting acceptance and trust.	"TechForLegal should be banned from interacting in the penal system as it is incapable of 100% fair judgment."

Prioritize Distribution Privacy	Participants showed a prioty in protecting distribution privacy.	“If TechForLegal chose "The AI does not favor brunettes, and it has a high chance to reveal the number of women in the data," the potential issue is that this aggregate information, while seemingly innocuous, could be used to infer patterns or biases within the AI’s training data related to gender. This could lead to scrutiny and potentially undermine trust in the AI’s fairness, even if the brunette bias is eliminated. While less directly harmful to an individual prisoner than revealing wealth, protecting sensitive information about the training data’s composition is important for maintaining the integrity and perceived impartiality of the system.”
Accept the tradeoff	Participant accept the given tradeoff, this code is for all the eight tradeoff	“Revealing the number of women in the data may not have an adverse impact compared to discriminating a group due to their hair type.”
Protect from manipulation, accept bias	Participants want protection from manipulation risk by accepting bias.	“If employees can manipulate the AI it’s dangerous for everyone.”
Manipulation leads to corruption	Manipulation of AI systems can enable corrupt practices.	“Easy manipulation could let biased staff unjustly increase sentences.”
Manipulation leads to discrimination	Manipulating an AI system can create discriminatory outcomes or unfair advantages.	“The different approaches all suggest manipulation which is never fair for anyone.”
Less bias, accept manipulation risk	Participants want less biased outcome by accepting the risk of manipulation.	“Brunettes potential being treated better by the system is a potential issue.”
Negative social outcome due to manipulation	Manipulating an AI system can produce negative social consequences, harming individuals or communities.	“If the AI were easy to manipulate, violent or dangerous criminals could be released back into the public.”
Loosing trust due to manipulation risk	Manipulation risks can cause people to lose trust in the AI system and its decisions.	“They could be seen as losing control of the company and loose trust.”
Balancing robustness and fairness	Participants chose an option in a way that balance robustness and fairness	“You don’t want it to be manipulated, or favor anyone, this is the best option.”

References

[1] Muhammad Naeem, Wilson Ozuem, Kerry Howell, and Silvia Ranfagni. A step-by-step process of thematic analysis to develop a conceptual model in qualitative research. *International journal of qualitative methods*, 22:16094069231205789, 2023.