A Game-Theoretic Analysis of Auditing Differentially Private Algorithms with Epistemically Disparate Herd

Ya-Ting Yang, Tao Zhang, and Quanyan Zhu

Abstract—Privacy-preserving AI algorithms are widely adopted in various domains, but the lack of transparency might pose accountability issues. While auditing algorithms can address this issue, machine-based audit approaches are often costly and time-consuming. Herd audit, on the other hand, offers an alternative solution by harnessing collective intelligence. Nevertheless, the presence of epistemic disparity among auditors, resulting in varying levels of expertise and access to knowledge, may impact audit performance. An effective herd audit will establish a credible accountability threat for algorithm developers, incentivizing them to uphold their claims. In this study, our objective is to develop a systematic framework that examines the impact of herd audits on algorithm developers using the Stackelberg game approach. The optimal strategy for auditors emphasizes the importance of easy access to relevant information, as it increases the auditors' confidence in the audit process. Similarly, the optimal choice for developers suggests that herd audit is viable when auditors face lower costs in acquiring knowledge. By enhancing transparency and accountability, herd audit contributes to the responsible development of privacy-preserving algorithms.

I. INTRODUCTION

AI and algorithmic decision-making have become pervasive in both business and society. However, when algorithms are treated as "black boxes" and their inner workings remain undisclosed, it becomes difficult to ensure that they perform as intended and adhere to necessary standards [1]. One specific category of algorithms that exemplifies this challenge is privacy-preserving algorithms [2]. For instance, platforms like Facebook Ad Recommendation Systems, Google SQL, and Safari have integrated differential privacy into their products to provide privacy protection. Nevertheless, verifying such claims can be arduous and intricate, for example, see [3]–[5].

Herd Audit: Auditing algorithms [6], [7] play a crucial role in tackling this challenge. However, traditional machine-based audit methods like direct scraping, sock puppet, and carrier puppet often necessitate the development of custom computer programs to gather data. Not only can these approaches be expensive, but they also consume a significant amount of time. A cost-effective alternative approach to auditing involves leveraging citizen science and crowd-sourcing principles to establish a democratic audit process that engages a diverse population of end users [8]. This concept gives rise to *herd-audit* (or group-audit) approaches. By empowering end users as auditors, we can foster a more democratic approach to algorithmic auditing while minimizing costs and time investments.

The Authors are with the Department of Electrical and Computer Engineering, New York University, Brooklyn, NY, 11201, USA; E-mail: {yy4348, tz636, qz494}@nyu.edu.
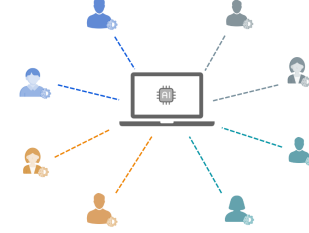
Fig. 1: A herd of diverse end-users act as auditors to inspect the AI algorithm used in the developed product.

Epistemic Disparity: One significant challenge in implementing herd-audit approaches is the presence of epistemic disparity [9], [10]. Not all users possess the same level of expertise or information required to conduct comprehensive audits of algorithms. A user-auditor with limited cognitive resources may inadvertently provide opportunities for algorithm developers to evade their responsibility. To some extent, incorporating audit into the algorithm design process itself establishes an accountability mechanism for developers. This accountability mechanism acts as an incentive for algorithm developers to uphold their claims and create responsible algorithms.

Game-Theoretic Framework: To design an effective herd-audit mechanism, this work aims to develop a comprehensive system framework that investigates the influence of herd-audit on algorithm developers. To accomplish the goal, the system framework adopts a Stackelberg game approach [11], [12]. In this approach, the developer assumes the role of the leader and determines the desired level of performance for differential privacy. The followers, comprising idiosyncratic end-users or auditors, are selected from a user population characterized by varying levels of epistemic capabilities. The proposed framework assumes that algorithms and their associated guarantees are clearly communicated to the end-users through a privacy protection agreement. This leader-and-follower structure allows us to analyze the optimal strategies employed by both the developer and the auditors, providing insights into the potential noncompliant behaviors of developers.

In order to capture the epistemic disparity experienced by end-users (auditors), this work employs a rational inattention model [13], [14], which takes into account the costs associated with accessing information during the decision-making process. We analyze the epistemic disparity among auditors, characterized by the epistemic factor, which measures the difficulty of accessing information. We find that auditors with lower epistemic factors exhibit higher audit confidence, indicating a better audit performance. Furthermore, our investigation reveals that a herd audit is a viable approach when auditors face lower costs in accessing information. In such circumstances, the algorithm developer is less likely to deviate

significantly from their claims. Our findings highlight the importance of reducing epistemic injustice as well as lowering information costs to enhance the effectiveness of herd audits. By doing so, we can foster a more reliable and accountable environment for the development of algorithms.

Related Works Algorithm auditing refers to the process of evaluating the algorithms used in systems or applications to ensure they are fair, transparent, unbiased, and comply with ethical standards [6]. In differential privacy, several machine-based verification methods have been proposed [3]–[5]. While there has been a rich literature on citizen science and its applications in crowdsensing [15], crowdsourcing [16], and crowd defense [17], herd audit is a concept in its infancy. It reduces auditing costs and poses a threat to developers, as public perception [18] can be influenced by the audit results.

The disparity in the capability of herd behaviors has been studied in collective intelligence [8], [19]–[21]. The literature has examined the performance [22], reliability [23], and trustworthiness [24] of participants engaged in outsourced tasks. Processes such as risk and reputation management [25], [26] have been utilized to understand the differences among participants. Numerous studies have focused on different cognitive behaviors in humans, including cognitive-behavioral theory [27]–[29] which elucidates how thoughts, beliefs, and cognitive processes shape behavior, and the theory of mind [30] that attributes mental states such as beliefs and emotions to predict individuals' behavior. In our work, we employ the concept of rational inattention, as studied in [31], which provides a framework that analyzes how decision-makers acquire information while considering associated costs, enabling investigations into cognitive impacts on audit decisions.

A game-theoretic approach is commonly employed to capture the threat posed by followers in dynamic games, such as ultimatum games [32], Stackelberg games [33], bargaining games [34], as well as contract [35], [36] and incentive mechanisms designs [37], [38]. Recently, there has been increased interest in the investigation of evasion behaviors [39]. This includes exploiting evasion-aware detection methods [40] and developing evaders for subsequent tests of collaborative cognition-assisted detector [41].

II. Herd Auditors with Epistemic Disparity

In the context of herd-auditing an algorithm, the auditor is uncertain about the true state $\omega \in \Omega = \{g, b\}$, where g indicates the null hypothesis, implying that the algorithm is consistent with the claim, while b is for the alternative hypothesis, meaning that the algorithm does not comply. The prior belief of state ω can be denoted as $\mu(\omega)$, implying the auditor's uncertainty in the algorithm's compliance.

In order to reduce the uncertainty, the auditor can obtain information s about the state according to the information-obtaining strategy $d(s|\omega)$. More specifically, s can be viewed as the outcome of the algorithm, and $d(s|\omega)$ indicates how the auditor accesses (obtains) it. The information s together with the obtaining strategy leads to a posterior belief of the state $\mu(\omega|s) = \frac{\mu(\omega)d(s|\omega)}{\sum_\omega \mu(\omega)d(s|\omega)}$.

Based on the information s (correspondingly, the posterior belief $\mu(\omega|s)$), the auditor can select an element from a
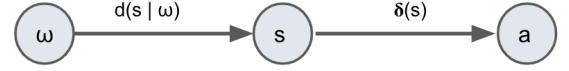

Fig. 2: An illustration of how the auditor performs audit. The auditor acquires information s about the unknown state ω using strategy d and then makes a decision a using strategy δ.

finite action set $a \in \mathcal{A} = \{T, F\}$, where T means reporting algorithm compliance, while F indicates reporting non-compliance. The decision rule $\delta : \mathcal{S} \mapsto \mathcal{A}$ aims to maximize the expected utility of $u(\omega, a)$, where $u : \Omega \times \mathcal{A} \mapsto \mathbb{R}$ is the utility of choosing action a when the state is ω.

However, the acquisition of information can incur costs, which can be viewed as the discrepancy between the prior belief $\mu(\omega)$ and the posterior belief $\mu(\omega|s)$ regarding the state ω. In conventional rational inattention research, a common method to model the cost is through the lens of Shannon mutual information. Furthermore, due to variations in epistemic disparities, the cost incurred for accessing information (i.e., reduction in uncertainty) differs among auditors. To account for this, we introduce the concept of an *epistemic factor* for each auditor, denoted as λ, which quantifies the differences in the cost experienced by different auditors when reducing the same amount of uncertainty. The larger value of λ implies harder access to relevant information, as the cost for the same amount of uncertainty reduction becomes higher. To this end, the auditor's objective becomes

$$\max_{d,\delta} \mathbb{E}[u(\omega, a)] - \lambda I(\omega; s), \tag{1}$$

where the expected utility is given by

$$\mathbb{E}[u(\omega, a)] = \sum_\omega \sum_a \mu(\omega)u(\omega, a) \sum_{s:\delta(s)=a} d(s|\omega), \tag{2}$$

and the information cost is expressed as

$$I(\omega; s) = \sum_\omega \sum_s d(s|\omega)\mu(\omega) \ln \frac{d(s|\omega)}{\sum_\omega d(s|\omega)\mu(\omega)}. \tag{3}$$

A. Bayes hypothesis testing as the auditor's decision rule

Conventionally, Bayes hypothesis testing deals with the optimization problem

$$\max_\delta \mathbb{E}[u(\omega, a)] = \sum_\omega \sum_a \mu(\omega)u(\omega, a) \sum_{\delta(s)=a} d(s|\omega) \tag{4}$$

with given distributions for both hypotheses $d(s|g)$ and $d(s|b)$ during decision-making, which coincides with the first term in the auditor's objective (1). According to detailed derivation in Appendix A, the optimal decision rule can be written as

$$\delta^*(s) = \begin{cases} T, & \frac{\mu(b)d(s|b)}{\mu(g)d(s|g)} < \frac{u(g,T)-u(g,F)}{u(b,F)-u(b,T)}, \\ F, & \frac{\mu(b)d(s|b)}{\mu(g)d(s|g)} > \frac{u(g,T)-u(g,F)}{u(b,F)-u(b,T)}, \\ \{T,F\}, & \frac{\mu(b)d(s|b)}{\mu(g)d(s|g)} = \frac{u(g,T)-u(g,F)}{u(b,F)-u(b,T)}, \end{cases} \tag{5}$$

which leads us to a threshold decision rule and can be viewed as making a decision based on the posteriors. We represent the optimal decision rule with given $d(s|g)$ and $d(s|b)$ as $\delta_d^*(s)$, and denote the information set partitioned by $\delta_d^*(s)$ as

$$\begin{cases} S_{d,T} = \{s : \delta_d^*(s) = T\}, \\ S_{d,F} = \{s : \delta_d^*(s) = F\}. \end{cases} \tag{6}$$

B. Auditor's choice of the information strategy

With the optimal decision rule δ_d^*, the auditor's objective:

$$\max_{d,\delta} \mathbb{E}[u(\omega,a)] - \lambda I(\omega;s), \text{ with } \delta = \delta_d^*, \quad (7)$$

which leads to the constrained optimization problem

$$\max_d \sum_\omega \sum_a \mu(\omega)u(\omega,a) \sum_{s:\delta_d^*(s)=a} d(s|\omega)$$
$$- \lambda \sum_\omega \sum_s d(s|\omega)\mu(\omega) \ln \frac{d(s|\omega)}{\sum_\omega d(s|\omega)\mu(\omega)}, \quad (8)$$
$$\text{s.t. } \sum_s d(s|\omega) = 1, d(s|\omega) \geq 0, \forall s \in \mathcal{S}, \forall \omega \in \Omega.$$

With detailed derivations in Appendix B we arrive at:

$$d(s|g) = \begin{cases} \frac{v(s)\exp(\frac{u(g,T)}{\lambda})}{y'(g)}, & s \in S_{d,T}, \\ \frac{v(s)\exp(\frac{u(g,F)}{\lambda})}{y'(g)}, & s \in S_{d,F}, \end{cases} \quad (9)$$

$$d(s|b) = \begin{cases} \frac{v(s)\exp(\frac{u(b,T)}{\lambda})}{y'(b)}, & s \in S_{d,T}, \\ \frac{v(s)\exp(\frac{u(b,F)}{\lambda})}{y'(b)}, & s \in S_{d,F}. \end{cases} \quad (10)$$

The corresponding posterior belief $\mu(g|s) = \frac{\mu(g)d(s|g)}{\sum_\omega \mu(\omega)d(s|\omega)} = \frac{\mu(g)d(s|g)}{v(s)}$ can then be written as

$$\mu(g|s) = \begin{cases} \frac{\mu(g)\exp(u(g,T)/\lambda)}{y'(g)}, & s \in S_{d,T}, \\ \frac{\mu(g)\exp(u(g,F)/\lambda)}{y'(g)}, & s \in S_{d,F}, \end{cases} \quad (11)$$

Note that the $s \in S_{d,T}$ case can be viewed as the posterior belief $\mu(g|s)$ given s that results in an action $a = T$ (i.e., $\mu(g|s) = \mu(g|T), \ s \in S_{d,T}$), while the $s \in S_{d,F}$ case can be viewed as the posterior belief $\mu(g|s)$ given s that results in an action $a = F$ (i.e., $\mu(g|s) = \mu(g|F), \ s \in S_{d,F}$). A similar expression can be found for $\mu(b|s)$.

$$\mu(b|s) = \begin{cases} \frac{\mu(b)\exp(u(b,T)/\lambda)}{y'(b)}, & s \in S_{d,T}, \\ \frac{\mu(b)\exp(u(b,F)/\lambda)}{y'(b)}, & s \in S_{d,F}, \end{cases} \quad (12)$$

where $y'(g)$ and $y'(b)$ are corresponding normalization terms.

Remark 1. *For an auditor with epistemic factor λ, the information-obtaining strategy represented by the conditional probability $d(s|\omega)$ is chosen if its resulting posterior belief $\mu(\omega|s)$ maximizes the value of $\mathbb{E}[u(\omega,a)] - \lambda I(\omega;s)$.*

The $\mu(g|s), \forall s \in S_{d,T}$, can also be interpreted as the *audit confidence* for making the decision $a = T$ when observing the information s. Since $u(g,T) > u(g,F)$, it is evident that auditors with a smaller epistemic factor λ have higher confidence in the audit process. This implies that auditors who can easily access relative information are more likely to perform better in the audit.

III. STACKELBERG HERD AUDIT GAME

To examine the impact of herd audit on the developer's incentive to behave irresponsibly, we formulate the interplay between the herd auditor (she) and the algorithm developer (he) as a Stackelberg herd audit game, depicted in Fig. 3.
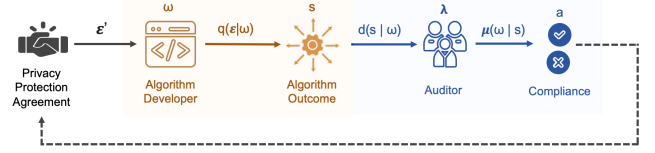

Fig. 3: The Stackelberg herd audit game.

A. Connection to differential privacy

We begin with the definition of ϵ-differential privacy.

Definition 1 (ϵ-DP). *A (randomized) mechanism $M : \mathcal{D} \mapsto \mathcal{B}$ is ϵ-differentially private (ϵ-DP) if for every pair of neighboring inputs $D_1, D_2 \in \mathcal{D}$, and for every (measurable) output set $B \in \mathcal{B}$, the probabilities of events $M(D_1; F, \epsilon) \in B$ and $M(D_2; F, \epsilon) \in B$ are closer than a factor of e^ϵ:*

$$Pr(M(D_1; F, \epsilon) \in B) \leq e^\epsilon \cdot Pr(M(D_2; F, \epsilon) \in B). \quad (13)$$

In the context of differential privacy, consider a scenario in which there is a public-known privacy protection agreement that requires ϵ' privacy budget. However, since more privacy budget (which means decreasing the privacy protection and making the results more distinguishable) often leads to better algorithm accuracy, the algorithm developer has the incentive to use some $\epsilon > \epsilon'$ when performing the algorithm, which creates irresponsibility. Hence, we consider the state $\omega = g$ means $\epsilon = \epsilon'$ and the state $\omega = b$ means $\epsilon > \epsilon'$. Since privacy protection is often achieved by adding noise, it is assumed that for an algorithm M with input dataset D, the privacy budget ϵ results in an output distribution $p(M(D)|\epsilon)$ for later usage.

B. Problem Setting for the Developer

Consider two types of algorithm developers g and b, and they play a mixed strategy for executing ϵ, which are $q(\epsilon|g)$ and $q(\epsilon|b)$, respectively (for discrete choices of $\epsilon \in \mathcal{E}$). Each ϵ results in an algorithm accuracy $A(\epsilon)$, where $A : \mathcal{E} \mapsto \mathbb{R}$, under the assumption that a larger ϵ leads to better accuracy.

Assumption 1. *Given algorithm M and input set D, a privacy budget ϵ leads to a unique output distribution $p(M(D)|\epsilon)$.*

Assumption 2. *For a given algorithm, the algorithmic accuracy under the privacy budget $\epsilon \in \mathcal{E} \subseteq \mathbb{R}$ is governed by $A : \mathcal{E} \mapsto \mathbb{R}$, and it is increasing in $\epsilon \in \mathcal{E}$.*

In this context, the developer's strategy $q(\epsilon|\omega)$ given his type ω will lead to the distributions for the two hypotheses

$$Q_g(s) = \sum_\epsilon p(s|\epsilon)q(\epsilon|g), \quad (14)$$
$$Q_b(s) = \sum_\epsilon p(s|\epsilon)q(\epsilon|b), \quad (15)$$

where $p(s|\epsilon)$ is the distribution $p(M(D)|\epsilon)$ in Assumption 1.

1) Responsible developer: For a responsible algorithm developer, the mixed strategy $q(\epsilon|g)$ should have mass 0 for $\epsilon > \epsilon'$, which means that he always provides privacy protection at least complies with the agreement. Moreover, in order to maximize $A(\epsilon)$, a responsible algorithm developer tends to put all the mass on $\epsilon = \epsilon'$ since $A(\epsilon) < A(\epsilon'), \forall \epsilon < \epsilon'$.

Proposition 1 (Responsible Developer's Strategy). *A responsible developer's mixed strategy reduces to a pure strategy by letting all the mass on $\epsilon = \epsilon'$. Hence, $Q_g(s) = \sum_\epsilon p(s|\epsilon)q(\epsilon|g) = p(s|\epsilon')$.*

2) Irresponsible developer: However, it is important to consider various scenarios involving an irresponsible algorithm developer who prioritizes algorithm performance and disregards compliance with the agreement. If there is no auditor or no penalty imposed when the developer fails to pass the audit (i.e., when the auditor determines that $a = F$), the irresponsible developer can choose an extremely large value for ϵ. Consequently, it is reasonable to assume that a penalty will be enforced if the irresponsible developer is detected. In such a situation, the irresponsible developer may attempt to maximize the probability of avoiding penalties, which is the probability of the auditor deciding $a = T$.

Assumption 3. *The irresponsible algorithm developer's mixed strategy will not put any mass on $\epsilon = \epsilon'$. That is, $q(\epsilon'|b) = 0$.*

C. Revisiting the Auditor's Problem

Considering that the penalty term for the irresponsible developer is influenced by the actions of the auditor, in terms of whether the irresponsible developer is caught or not, it is necessary to reexamine the problem from the auditor's perspective when the developer is also a strategic player aiming to evade the audit. We reformulate the auditor's problem by letting $u(g, T) = u(b, F) = 0$ and setting the penalty terms $u(b, T)$ and $u(g, F)$ to negative values. However, within the context of DP, it is important to note that the distributions for these hypotheses are predefined by the output distribution $p(s|\epsilon)$ and the developer's mixed strategy given the observed information s. The audit confidences are analogous to those provided in (11) and (12).

Assumption 4. *Assume that $u(b, T) < 0$ and $u(g, F) < 0$ are the negative utilities for making wrong audit decisions.*

Given the distributions for the two hypotheses $Q_g(s)$ and $Q_b(s)$, the auditor aims to achieve the following:

$$
\begin{aligned}
\max_r \quad & u(g, F) \sum_s \left[\sum_\epsilon p(s|\epsilon)q(\epsilon|g) \right] r(b|s) \\
& + u(b, T) \sum_s \left[\sum_\epsilon p(s|\epsilon)q(\epsilon|b) \right] r(g|s) \\
& - \lambda \, \mathbb{E}_s \left[D_{kL}(r(\omega|s) \| \mu(\omega)) \right],
\end{aligned}
$$
(16)

$$
\text{s.t.} \sum_\omega r(\omega|s) = 1, r(w|s) \geq 0, \forall w \in \{g, b\}, \forall s \in \mathcal{S},
$$

where the first two terms put negative weights on the audit error, and the last term quantifies the expected reduction in uncertainty for the state ω, measured in terms of the Kullback–Leibler (KL) divergence:

$$
\lambda \sum_s \left[Q_g(s) + Q_b(s) \right] \sum_\omega r(\omega|s) \log \frac{r(\omega|s)}{\mu(\omega)}.
$$

The decision of $r(g|s)$, $r(b|s)$ already incorporate the auditor's information strategy $d(s|\omega)$ since $r(\omega|s) = \frac{\mu(\omega)d(s|\omega)}{\sum_\omega \mu(\omega)d(s|\omega)}$.

D. Revisit the Irresponsible Developer's Problem

Until now, the irresponsible developer's objective has become the following.

$$
\max_{q(\cdot|b)} \sum_\epsilon q(\epsilon|b)A(\epsilon) + \beta \sum_s \left[\sum_\epsilon p(s|\epsilon)q(\epsilon|b) \right] r(g|s),
$$
(17)

with $r(g|s)$ comes from the auditor's problem. The former term is the expected algorithm accuracy, and the latter term corresponds to the false negative rate of the auditor's decision, which is the rate of the irresponsible developer successfully passing the audit (and thus, the irresponsible developer seeks to maximize it). Note that $\beta > 0$ indicates the irresponsible developer's preference for the two goals.

IV. Equilibrium Analysis

For illustrative purposes, we work through an example where $|\mathcal{E}| = 3$ in Appendix C. Besides, we assume that the distinguishability—quantified by distance measures such as the Kullback–Leibler divergence—between the output distributions $p(\cdot|\epsilon)$ and $p(\cdot|\epsilon')$ increases when the difference between ϵ and ϵ' expands.

A. The auditor's optimal strategy

With the example in Appendix C and derivations in Appendix D, the auditor's $r(g|s)$ and $r(b|s)$ that optimally solves problem (16) can be written as:

$$
r(g|s) = \mu(g) \exp\left(u(b, T)Q_b(s)/\lambda v(s) \right) / y'(s), \quad (18)
$$

$$
r(b|s) = \mu(b) \exp\left(u(g, F)Q_g(s)/\lambda v(s) \right) / y'(s), \quad (19)
$$

and $y'(s) = \mu(g) \exp\left(\frac{u(b,T)Q_b(s)}{\lambda v(s)} \right) + \mu(b) \exp\left(\frac{u(g,F)Q_g(s)}{\lambda v(s)} \right)$ denotes the normalization term. We can observe that $r(g|s)$ and $r(b|s)$ share a similar form as (11) and (12).

Proposition 2. *The strategy specified by (18) and (19) is optimal for the auditor with epistemic factor λ.*

Remark 2. *The results coincide with the intuition. We first take a look at $r(g|s)$. Recall that $\frac{u(b,T)}{\lambda}$ is negative. If the penalty term $u(b, T)$ is the same across all the auditors, the auditor with a larger epistemic factor λ achieves $r(g|s)$ that is closer to $\mu(g)$. Combining with the auditor's objective in the maximization problem (16), it means that the larger-λ auditor might have a larger false negative error. Similarly, for $r(b|s)$, the larger-λ auditor might have a larger false positive error.*

B. The irresponsible developer's optimal strategy

The irresponsible developer's problem (17) is organized into

$$
\begin{aligned}
& \sum_\epsilon q(\epsilon|b)A(\epsilon) + \beta \sum_s \left[\sum_\epsilon p(s|\epsilon)q(\epsilon|b) \right] r(g|s) \\
& = \sum_{\epsilon_i \in \mathcal{E}} q(\epsilon_i|b) \left[A(\epsilon_i) + \beta \sum_s r(g|s)p(s|\epsilon_i) \right].
\end{aligned}
$$
(20)

The irresponsible developer determines his optimal pure strategy ϵ to maximize (20). Specifically, the irresponsible developer assigns $q(\epsilon|b) = 1$ to the ϵ that achieves the largest

$[A(\epsilon) + \beta \sum_s r(g|s)p(s|\epsilon)]$. This leads us to the following propositions and remarks.

Proposition 3. *The irresponsible developer's optimal strategy is choosing the ϵ that maximizes $[A(\epsilon) + \beta \sum_s r(g|s)p(s|\epsilon)]$.*

Proposition 4. *If algorithm accuracy $A(\epsilon)$ is increasing in ϵ, the irresponsible developer always chooses the largest ϵ if $r(g|s) = r(g|s'), \forall s, s' \in \mathcal{S}$.*

Proof. We sketch the proof in Appendix E. □

Remark 3. *The irresponsible developer violates as much as possible when the epistemic factor for the auditor $\lambda = \infty$.*

Remark 4. *When the auditor's epistemic factor λ is small, indicating easy access to relevant information, an irresponsible developer is more likely to violate the claim.*

Remark 5. *If the auditor's epistemic factor λ is large, it is likely that an irresponsible developer with a larger β (placing more value on the success rate of passing audits) will also tend to violate the claim more severely.*

C. Auditor's audit confidence and epistemic factor

With respect to Fig. 4, the optimal solution to the auditor's problem given by (18) and (19) establishes a relationship between the epistemic factor λ and the auditor's confidence $r(\cdot|s)$ under fixed utilities $u(\omega, a)$. Let $\chi(s) = [u(g, F)Q_g(s) - u(b, T)Q_b(s)]/v(s)$ and $\phi(s) = [u(g, F)Q_g(s) + u(b, T)Q_b(s)]/(\lambda v(s))$. Taking the partial derivative of $r(g|s)$ with respect to λ yields $\frac{\partial r(g|s)}{\partial \lambda} = \frac{\mu(g)\mu(b)\chi(s)\exp(\phi(s))}{\lambda^2 y'(s)^2}$. Here, if the developer is irresponsible, then he never chooses a privacy budget ϵ that is equal to the claimed budget ϵ_0. Hence, $\chi(s) \neq 0$. The term $\partial r(g|s)/\partial \lambda$ is (strictly) positive if $\chi(s) > 0$ and (strictly) negative otherwise. When $\chi(s) < 0$, $r(g|s)$ is close to 1 when λ goes close to 0. Furthermore, the audit confidences for g and b become close to 0.5 when λ increases, which reveals that higher λ leads to a weaker incentive to acquire more accurate information, thereby inducing lower audit confidences.

Similarly, $\frac{\partial r(b|s)}{\partial \lambda} = \frac{-\mu(g)\mu(b)\chi(s)\exp(\phi(s))}{\lambda^2 y'(s)^2}$. The term $\partial r(b|s)/\partial \lambda$ is positive if $-\chi(s) > 0$ and negative otherwise. When $-\chi(s) < 0$, $r(b|s)$ is close to 1 when λ goes close to 0. Furthermore, the audit confidences for g and b become closer to 0.5 when λ increases, which coincides with the setting that higher λ leads to a weaker incentive to acquire more accurate information, thereby inducing lower audit confidences. Note that audit confidence is determined by optimizing the objective, which consists of penalties for audit errors and costs associated with information acquisition. In this context, it is important to carefully select reasonable intervals for $u(\omega, a)$ and λ. In practice, as auditors are end-users for the algorithm, and given the disparities in end-users across different algorithms, the range for the epistemic factor needs to be contingent upon the ease with which corresponding end-users of the algorithm can access relevant information.

D. Irresponsible developer's choice and auditor's confidence

According to (14) and (15), the irresponsible developer's budget choice determines $Q_b(\cdot)$ given $p(\cdot)$. Hence, (18) and
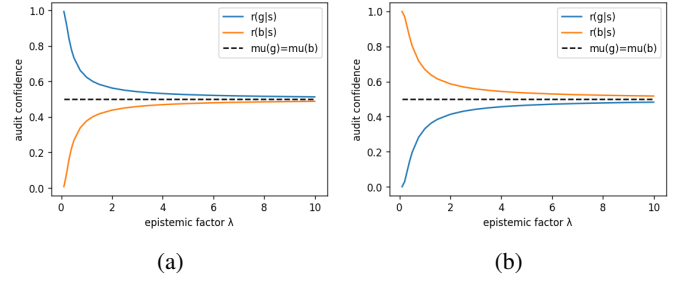
Fig. 4: Trends between the auditor's confidence (posterior belief) and the auditor's epistemic factor ($\mu(g) = \mu(b) = 0.5, u(b, T) = u(g, F) = -1$). Left: $\chi(s) < 0$ ($Q_b(s)/v(s) = 0.25$). Right: $-\chi(s) < 0$ ($Q_b(s)/v(s) = 0.85$).

(19) (shown in Fig. 5) also establish a relationship between the irresponsible developer's choice and the auditor's confidence.
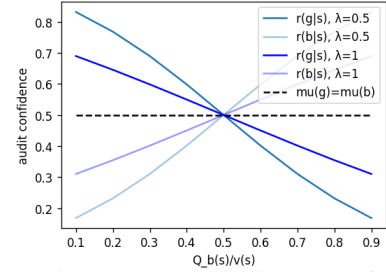
Fig. 5: The trend between the auditor's confidence and the irresponsible developer's choice that leads to $Q_b(s)/v(s)$. Here, $\mu(g) = \mu(b) = 0.5, u(b, T) = u(g, F) = -1, \lambda = 1$.

By taking partial derivative of $r(g|s)$ with respect to $\frac{Q_b(s)}{v(s)}$, we obtain $\frac{\partial r(g|s)}{\partial Q_b(s)/v(s)} = \frac{\mu(g)\mu(b)(u(g,F)+u(b,T))\exp(\phi(s))}{\lambda y'(s)^2}$, which is negative since $u(g, F) + u(b, T) < 0$. Additionally, as the value of λ increases (when auditors incur higher costs for information acquisition), the magnitude of $\partial r(g|s)/\partial(\frac{Q_b(s)}{v(s)})$ decreases, implying relatively less influence on audit confidence. This trend is evident in Figure 5, where a greater λ corresponds to a flatter curve for $r(g|s)$.

V. DISCUSSION AND CONCLUSIONS

Herd audit is a collective mechanism that empowers users to hold algorithm developers accountable, fostering the development of compliant and responsible digital products for the betterment of society. In this study, we examine herd audit through a game-theoretic lens, capturing the interactions between an idiosyncratic user and a privacy-preserving algorithm developer. Our framework adopts a Stackelberg game approach, enabling us to assess the impact of herd audit on responsible algorithm design and understand selfish and irresponsible strategies in worst-case scenarios.

We have specifically explored the presence of auditors with varying cognitive and reasoning capabilities, capturing epistemic disparities. Within our game-theoretic framework, we have consolidated the concept of rational inattention. The optimal strategy for auditors underscores the importance of easy access to relevant information, which enhances their confidence in the herd-audit process. Similarly, the optimal decision for algorithm developers has revealed that herd audit is a viable approach when auditors face lower costs in accessing knowledge, as denoted by smaller epistemic factors.

Based on our findings, we conclude that herd audit poses a credit threat to developers and plays a vital role in promoting the responsible development of privacy-preserving algorithms. In future work, we aim to enrich the game-theoretic framework by incorporating end-users' incentives. This extension allows us to design an incentive mechanism that encourages participation in herd audits. Additionally, we plan to explore the fusion of distributed audits alongside a central audit center. Leveraging tools from decentralized hypothesis testing, game theory, information theory, and differential privacy, this research direction holds promise for advancing the field further.

REFERENCES

[1] J. Guszcza, I. Rahwan, W. Bible, M. Cebrian, and V. Katyal, "Why we need to audit algorithms," 2018. [Online]. Available: https://hdl.handle.net/21.11116/0000-0003-1C9E-D

[2] C. Dwork, "Differential privacy: A survey of results," in *Theory and Applications of Models of Computation: 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008. Proceedings 5.* Springer, 2008, pp. 1–19.

[3] Z. Ding, Y. Wang, G. Wang, D. Zhang, and D. Kifer, "Detecting violations of differential privacy," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018, pp. 475–489.

[4] B. Bichsel, T. Gehr, D. Drachsler-Cohen, P. Tsankov, and M. Vechev, "Dp-finder: Finding differential privacy violations by sampling and optimization," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 508–524.

[5] Y. Han and S. Martínez, "A numerical verification framework for differential privacy in estimation," *IEEE Control Systems Letters*, vol. 6, pp. 1712–1717, 2021.

[6] J. Bandy, "Problematic machine behavior: A systematic literature review of algorithm audits," *Proceedings of the acm on human-computer interaction*, vol. 5, no. CSCW1, pp. 1–34, 2021.

[7] B. Mittelstadt, "Automation, algorithms, and politics— auditing for transparency in content personalization systems," *International Journal of Communication*, vol. 10, p. 12, 2016.

[8] J. M. Leimeister, "Collective intelligence," *Business & Information Systems Engineering*, vol. 2, pp. 245–248, 2010.

[9] M. Fricker, *Epistemic injustice: Power and the ethics of knowing.* Oxford University Press, 2007.

[10] H. Grasswick, "Epistemic injustice in science," in *The Routledge handbook of epistemic injustice.* Routledge, 2017, pp. 313–323.

[11] M. H. Manshaei, Q. Zhu, T. Alpcan, T. Başar, and J.-P. Hubaux, "Game theory meets network security and privacy," *ACM Computing Surveys (CSUR)*, vol. 45, no. 3, pp. 1–39, 2013.

[12] F. Fang, S. Liu, A. Basak, Q. Zhu, C. D. Kiekintveld, and C. A. Kamhoua, "Introduction to game theory," *Game Theory and Machine Learning for Cyber Security*, pp. 21–46, 2021.

[13] F. Matějka and A. McKay, "Rational inattention to discrete choices: A new foundation for the multinomial logit model," *American Economic Review*, vol. 105, no. 1, pp. 272–298, 2015.

[14] A. Caplin and M. Dean, "Revealed preference, rational inattention, and costly information acquisition," *American Economic Review*, vol. 105, no. 7, pp. 2183–2203, July 2015.

[15] F. Restuccia, N. Ghosh, S. Bhattacharjee, S. K. Das, and T. Melodia, "Quality of information in mobile crowdsensing: Survey and research challenges," *ACM Transactions on Sensor Networks (TOSN)*, vol. 13, no. 4, pp. 1–43, 2017.

[16] Y. Zhao and Q. Zhu, "Evaluation on crowdsourcing research: Current status and future direction," *Information systems frontiers*, vol. 16, pp. 417–434, 2014.

[17] J. Pawlick and Q. Zhu, "Active crowd defense," *Game Theory for Cyber Deception: From Theory to Applications*, pp. 147–167, 2021.

[18] H. Frye, "The technology of public shaming," *Social Philosophy and Policy*, vol. 38, no. 2, pp. 128–145, 2021.

[19] H. Yu, C. Miao, C. Leung, Y. Chen, S. Fauvel, V. R. Lesser, and Q. Yang, "Mitigating herding in hierarchical crowdsourcing networks," *Scientific reports*, vol. 6, no. 1, p. 4, 2016.

[20] I. Comeig, E. Mesa-Vázquez, P. Sendra-Pons, and A. Urbano, "Rational herding in reward-based crowdfunding: An mturk experiment," *Sustainability*, vol. 12, no. 23, p. 9827, 2020.

[21] C. Eickhoff, "Cognitive biases in crowdsourcing," in *Proceedings of the eleventh ACM international conference on web search and data mining*, 2018, pp. 162–170.

[22] R. R. Morris, M. Dontcheva, and E. M. Gerber, "Priming for better performance in microtask crowdsourcing environments," *IEEE Internet Computing*, vol. 16, no. 5, pp. 13–19, 2012.

[23] D. R. Karger, S. Oh, and D. Shah, "Budget-optimal task allocation for reliable crowdsourcing systems," *Operations Research*, vol. 62, no. 1, pp. 1–24, 2014.

[24] K. Wang, X. Qi, L. Shu, D.-j. Deng, and J. J. Rodrigues, "Toward trustworthy crowdsourcing in the social internet of things," *IEEE Wireless Communications*, vol. 23, no. 5, pp. 30–36, 2016.

[25] M. Allahbakhsh, A. Ignjatovic, B. Benatallah, E. Bertino, N. Foo *et al.*, "Reputation management in crowdsourcing systems," in *8th International conference on collaborative computing: networking, applications and worksharing (CollaborateCom).* IEEE, 2012, pp. 664–671.

[26] Y. Yu, S. Liu, L. Guo, P. L. Yeoh, B. Vucetic, and Y. Li, "Crowdr-fbc: A distributed fog-blockchains for mobile crowdsourcing reputation management," *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8722–8735, 2020.

[27] A. A. González-Prendes and S. M. Resko, "Cognitive-behavioral theory," 2012.

[28] D. Fum, F. Del Missier, A. Stocco *et al.*, "The cognitive modeling of human behavior: Why a model is (sometimes) better than 10,000 words," *Cognitive Systems Research*, vol. 8, no. 3, pp. 135–142, 2007.

[29] L. Huang and Q. Zhu, *Cognitive Security: A System-Scientific Approach.* Springer Nature, 2023.

[30] J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin, "An integrated theory of the mind." *Psychological review*, vol. 111, no. 4, p. 1036, 2004.

[31] C. A. Sims, "Implications of rational inattention," *Journal of monetary Economics*, vol. 50, no. 3, pp. 665–690, 2003.

[32] S. Rajtmajer, A. Squicciarini, J. M. Such, J. Semonsen, and A. Belmonte, "An ultimatum game model for the evolution of privacy in jointly managed content," in *Decision and Game Theory for Security: 8th International Conference, GameSec 2017, Vienna, Austria, October 23-25, 2017, Proceedings.* Springer, 2017, pp. 112–130.

[33] C. Casorrán, B. Fortz, M. Labbé, and F. Ordóñez, "A study of general and security stackelberg game formulations," *European journal of operational research*, vol. 278, no. 3, pp. 855–868, 2019.

[34] D. Guerrero, A. A. Carsteanu, and J. B. Clempner, "Solving stackelberg security markov games employing the bargaining nash approach: Convergence analysis," *Computers & Security*, vol. 74, pp. 240–257, 2018.

[35] J. Chen and Q. Zhu, "Optimal contract design under asymmetric information for cloud-enabled internet of controlled things," in *Decision and Game Theory for Security: 7th International Conference, GameSec 2016, New York, NY, USA, November 2-4, 2016, Proceedings.* Springer, 2016, pp. 329–348.

[36] R. Zhang and Q. Zhu, "Flipin: A game-theoretic cyber insurance framework for incentive-compatible cyber risk management of internet of things," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2026–2041, 2019.

[37] Q. Zhu, C. Fung, R. Boutaba, and T. Basar, "Guidex: A game-theoretic incentive-based mechanism for intrusion detection networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 11, pp. 2220–2230, 2012.

[38] L. Huang and Q. Zhu, "Duplicity games for deception design with an application to insider threat mitigation," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4843–4856, 2021.

[39] K. Horák, Q. Zhu, and B. Bošanský, "Manipulating adversary's belief: A dynamic game approach to deception by design for proactive network security," in *Decision and Game Theory for Security: 8th International Conference, GameSec 2017, Vienna, Austria, October 23-25, 2017, Proceedings.* Springer, 2017, pp. 273–294.

[40] Y. Hu and Q. Zhu, "Evasion-aware neyman-pearson detectors: A game-theoretic approach," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2022, pp. 6111–6117.

[41] S. N. Narayanan, A. Ganesan, K. Joshi, T. Oates, A. Joshi, and T. Finin, "Early detection of cybersecurity threats using collaborative cognition," in *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*, 2018, pp. 354–363.

APPENDIX

A. Proof of optimal decision rule $\delta^*(s)$

For the null hypothesis, $d(s|g)$, and the alternative hypothesis, $d(s|b)$, the expected utility for problem HT in (**??**) can be reformulated as

$$\mathbb{E}[u(\omega, a)] = \mu(g)u(g, F) + \mu(b)u(b, F)$$
$$+ \sum_{\delta(s)=T} \left\{ \mu(g)\big[u(g, T) - u(g, F)\big]d(s|g) \right.$$
$$\left. - \mu(b)\big[u(b, F) - u(b, T)\big]d(s|b) \right\}.$$

Therefore, to maximize the expected utility, the auditor must decide $\delta(s) = T$ if $\mu(g)\big[u(g, T) - u(g, F)\big]d(s|g) > \mu(b)\big[u(b, F) - u(b, T)\big]d(s|b)$. This completes the proof.

B. Proof of optimal information strategy d

To analyze the problem, we use the method of Lagrange multipliers and denote

$$J(d, y) = \sum_{\omega} \sum_{a} \mu(\omega)u(\omega, a) \sum_{s:\delta_d^*(s)=a} d(s|\omega)$$
$$- \lambda \sum_{\omega} \sum_{s} d(s|\omega)\mu(\omega) \ln \frac{d(s|\omega)}{\sum_{\omega} d(s|\omega)\mu(\omega)} - \sum_{\omega} y(\omega)d(s|\omega),$$

with the last term corresponding to the constraint that $d(s|\omega)$ should be a conditional probability mass function.

Then, for $d(s|g)$ with $s \in S_{d,T}$, according to the first-order and the second-order condition,

$$\frac{\partial J(d, y)}{\partial d(s|g)} = \mu(g)u(g, T) - \lambda\mu(g)\log(\frac{d(s|g)}{v(s)}) - y(g) = 0,$$
$$\frac{\partial^2 J(d, y)}{\partial d(s|g)^2} = -\lambda\mu(g)\frac{\mu(b)d(s|b)}{d(s|g)v(s)} \leq 0,$$

where $v(s) = \sum_{\omega} d(s|\omega)\mu(\omega)$. Letting $\log(y'(g)) = \frac{y(g)}{\lambda\mu(g)}$ leads to the following $d(s|g)$ that maximizes (8).

$$\lambda\mu(g)\left[\frac{u(g,T)}{\lambda} - \log(\frac{d(s|g)}{v(s)}) - \log y'(g)\right] = 0,$$
$$d(s|g) = \frac{v(s)\exp(\frac{u(g,T)}{\lambda})}{y'(g)}.$$

Note that $y'(g)$ is the normalization term. Similarly, for $d(s|g)$ with $s \in S_{d,F}$ and $d(s|b)$, we can get the information-obtaining strategy in (9) and (10)

C. An illustrative example for equilibrium analysis

We consider a scenario where the cardinality of the set $\mathcal{E}$ is three; i.e., $|\mathcal{E}| = 3$ with $\mathcal{E} = \{\epsilon_l, \epsilon_m, \epsilon_h\}$, where $\epsilon_l < \epsilon_m < \epsilon_h$ and it's assumed that the claimed differential privacy budget is $\epsilon' = \epsilon_l$. Then, the two hypotheses become

$$Q_g(s) = \sum_{\epsilon} p(s|\epsilon)q(\epsilon|g) = p(s|\epsilon') = p(s|\epsilon_l),$$

$$Q_b(s) = \sum_{\epsilon} p(s|\epsilon)q(\epsilon|b) = p(s|\epsilon_m)q(\epsilon_m|b) + p(s|\epsilon_h)q(\epsilon_h|b).$$

According to derivations in Appendix D, the strategy specified by (18) and (19) is optimal for the auditor with epistemic factor λ.

We then shift our focus to the irresponsible developer's strategy. The irresponsible developer endeavors to enhance algorithmic accuracy while concurrently maximizing the probability of evading detection by the auditor, thereby increasing the likelihood of being perceived as a responsible developer. Hence, the irresponsible developer's decision-making can be described by the following optimization problem:

$$\max_{q(\cdot|b)} \; \big[q(\epsilon_m|b)A(\epsilon_m) + q(\epsilon_h|b)A(\epsilon_h)\big]$$
$$+ \beta \sum_{s} \left[p(s|\epsilon_m)q(\epsilon_m|b) + p(s|\epsilon_h)q(\epsilon_h|b)\right]r(g|s). \tag{21}$$

By leveraging $q(\epsilon_m|b) = 1 - q(\epsilon_h|b)$, we rewrite the problem (21) as follows:

$$\max_{q(\epsilon_l|b)} \; A(\epsilon_h) + \beta \sum_{s} r(g|s)p(s|\epsilon_h) + \left\{ \big[A(\epsilon_m) - A(\epsilon_h)\big] \right.$$
$$\left. + \beta \sum_{s} r(g|s)\big[p(s|\epsilon_m) - p(s|\epsilon_h)\big] \right\}q(\epsilon_m|b). \tag{22}$$

Since the first two terms $A(\epsilon_h) + \beta \sum_{s} r(g|s)p(s|\epsilon_h)$ are independent of $q(\cdot|b)$, (22) suggests the following strategy for the irresponsible developer: let $\Delta A = A(\epsilon_m) - A(\epsilon_h)$,

$$\begin{cases} q(\epsilon_m|b) = 1, & \Delta A + \beta \sum_{s} r(g|s)\big[p(s|\epsilon_m) - p(s|\epsilon_h)\big] > 0 \\ q(\epsilon_h|b) = 1, & \text{otherwise.} \end{cases}$$

That is, the irresponsible developer has a pure strategy by choosing either $q(\epsilon_m|b) = 1$ or $q(\epsilon_h|b) = 1$.

D. Proof of auditor's strategy r

In (16), the KL divergence term with a negative sign is concave with respect to the decision variables $r(\cdot)$ given fixed priors $\mu(\cdot)$. Therefore, the combination of the terms in the objective function forms a weighted sum of concave functions. This makes the overall objective function concave. Given the linear constraints, the feasibility set is convex. Hence, the optimization problem (16) is a concave maximization over a convex set.

The Lagrangian corresponding to (16) is then given by

$$J(r, y, z) = u(g, F) \sum_{s} \left[\sum_{\epsilon} p(s|\epsilon)q(\epsilon|g) \right] r(b|s)$$
$$+ u(b, T) \sum_{s} \left[\sum_{\epsilon} p(s|\epsilon)q(\epsilon|b) \right] r(g|s)$$
$$- \lambda \sum_{s} \left[\sum_{\epsilon} p(s|\epsilon)q(\epsilon|g) + \sum_{\epsilon} p(s|\varepsilon)q(\epsilon|b) \right]$$
$$\sum_{\omega} r(\omega|s) \log \frac{r(\omega|s)}{\mu(\omega)} - yr(g|s) - zr(b|s), \tag{23}$$

where $y \in \mathbb{R}$ and $z \in \mathbb{R}$ are the associated Lagrange multipliers. Then, the first-order condition concerning $r(g|s)$ implies

$$
\begin{aligned}
\frac{\partial J(r)}{\partial r(g|s)} &= u(b,T)\Big[p(s|\epsilon_m)q(\epsilon_m|b) + p(s|\epsilon_h)q(\epsilon_h|b)\Big] \\
&\quad - \lambda\Big[p(s|\epsilon_l) + p(s|\epsilon_m)q(\epsilon_m|b) + p(s|\epsilon_h)q(\epsilon_h|b)\Big] \\
&\quad \Big(\log\frac{r(g|s)}{\mu(g)} + 1\Big) - y \\
&= 0.
\end{aligned}
$$

Hence, we obtain

$$
u(b,T)Q_b(s) - \lambda v(s)\Big(\log\frac{r(g|s)}{\mu(g)} + 1\Big) - y = 0,
$$

$$
\frac{u(b,T)Q_b(s)}{\lambda v(s)} - \Big(\frac{y}{\lambda v(s)} + 1\Big) = \log\frac{r(g|s)}{\mu(g)}.
$$

By letting $\log y'(s) = \big(\frac{y}{\lambda v(s)} + 1\big)$, $r(g|s)$ can then be written as (18). We can get $r(b|s)$ described in (19) with a similar process.

E. Proof of Proposition 4

We sketch the proof for $|\mathcal{S}| = 2$ with $\mathcal{S} = \{s_1, s_2\}$. In this example, $p(s_1|\epsilon_m) + p(s_2|\epsilon_m) = 1$ and $p(s_1|\epsilon_h) + p(s_2|\epsilon_h) = 1$, then $r(g|s_1)[p(s_1|\epsilon_m) - p(s_1|\epsilon_h)] + r(g|s_2)[p(s_2|\epsilon_m) - p(s_2|\epsilon_h)] = 1 - 1 = 0$ if $r(g|s_1) = r(g|s_2)$.

Hence, $A(\epsilon_m) - A(\epsilon_h) + \sum_s r(g|s)\big[p(s|\epsilon_m) - p(s|\epsilon_h)\big] < 0$ in the case where $A(\epsilon_m) < A(\epsilon_h)$, which leads to $q(\epsilon_m|b) = 0$ and $q(\epsilon_h|b) = 1$. This completes the proof.