Electronic Companion - "Beyond the Black Box: Unraveling the Role of Explainability in Human-AI Collaboration"

Tamer Boyacı
ESMT Berlin, 10178 Berlin, Germany, tamer.boyaci@esmt.org

Caner Canyakmaz
TBS Business School, Toulouse 31000, France, c.canyakmaz@tbs-education.fr

Francis de Véricourt
ESMT Berlin, 10178 Berlin, Germany, francis.devericourt@esmt.org

Appendix A: Proofs of Results

This section contains proofs of all results in the main paper and the appendices.

Proof of Lemma 1 Let $\xi \in (1/2, 1)$. Conditioning on machine quality, the posterior probability that the state is $\omega = 1$ given a positive prediction and a high-quality explanation is

$$\mu^{+\hbar} = \mathbb{P}\{\omega = 1 | s = +, e = \hbar, X = \mathfrak{l}\}\, \mathbb{P}\{X = \mathfrak{l} | s = +, e = \hbar\}$$
$$+ \mathbb{P}\{\omega = 1 | s = +, e = \hbar, X = \mathfrak{h}\}\, \mathbb{P}\{X = \mathfrak{h} | s = +, e = \hbar\} = \mu\left(1 - \theta^{+\hbar}\right) + \theta^{+\hbar}$$

where

$$\begin{aligned}
\theta^{+\hbar} &= \mathbb{P}\{X = \mathfrak{h} | s = +, e = \hbar\} \\
&= \frac{\mathbb{P}\{s = +, e = \hbar | X = \mathfrak{h}\}\, \mathbb{P}\{X = \mathfrak{h}\}}{\mathbb{P}\{s = +, e = \hbar | X = \mathfrak{h}\}\, \mathbb{P}\{X = \mathfrak{h}\} + \mathbb{P}\{s = +, e = \hbar | X = \mathfrak{l}\}\, \mathbb{P}\{X = \mathfrak{l}\}} \\
&= \frac{\mathbb{P}\{e = \hbar | X = \mathfrak{h}\}\, \mathbb{P}\{s = + | X = \mathfrak{h}\}\, \mathbb{P}\{X = \mathfrak{h}\}}{\mathbb{P}\{e = \hbar | X = \mathfrak{h}\}\, \mathbb{P}\{s = + | X = \mathfrak{h}\}\, \mathbb{P}\{X = \mathfrak{h}\} + \mathbb{P}\{e = \hbar | X = \mathfrak{l}\}\, \mathbb{P}\{s = + | X = \mathfrak{l}\}\, \mathbb{P}\{X = \mathfrak{l}\}} \\
&= \frac{\xi\mu\theta}{\xi\mu\theta + \frac{1}{2}(1-\xi)(1-\theta)}
\end{aligned}$$

is found by using

$$\mathbb{P}\{s = + | X = \mathfrak{h}\} = \mathbb{P}\{s = + | X = \mathfrak{h}, \omega = 1\}\, \mathbb{P}\{\omega = 1 | X = \mathfrak{h}\} + \mathbb{P}\{s = + | X = \mathfrak{h}, \omega = 0\}\, \mathbb{P}\{\omega = 0 | X = \mathfrak{h}\} = \mu,$$

$$\mathbb{P}\{s = + | X = \mathfrak{l}\} = \mathbb{P}\{s = + | X = \mathfrak{l}, \omega = 1\}\, \mathbb{P}\{\omega = 1 | X = \mathfrak{l}\}$$
$$+ \mathbb{P}\{s = + | X = \mathfrak{l}, \omega = 0\}\, \mathbb{P}\{\omega = 0 | X = \mathfrak{l}\} = \frac{1}{2}\mu + \frac{1}{2}(1-\mu) = \frac{1}{2},$$

and $\mathbb{P}\{e=\hbar, s=+|X=\mathfrak{h}\}=\mathbb{P}\{e=\hbar|X=\mathfrak{h}\}\mathbb{P}\{s=+|X=\mathfrak{h}\}$ due to the independence of s and e given x. In a similar fashion, we have

$$\mu^{+\ell}=\mu\left(1-\theta^{+\ell}\right)+\theta^{+\ell}$$
$$\mu^{-\hbar}=\mu\left(1-\theta^{-\hbar}\right)$$
$$\mu^{-\ell}=\mu\left(1-\theta^{-\ell}\right)$$

where

$$\theta^{+\ell}=\mathbb{P}\{X=\mathfrak{h}|s=+,e=\ell\}=\frac{(1-\xi)\mu\theta}{(1-\xi)\mu\theta+\frac{1}{2}\xi(1-\theta)}$$
$$\theta^{-\hbar}=\mathbb{P}\{X=\mathfrak{h}|s=-,e=\hbar\}=\frac{\xi(1-\mu)\theta}{\xi(1-\mu)\theta+\frac{1}{2}(1-\xi)(1-\theta)}$$
$$\theta^{-\ell}=\mathbb{P}\{X=\mathfrak{h}|s=-,e=\ell\}=\frac{(1-\xi)(1-\mu)\theta}{(1-\xi)(1-\mu)\theta+\frac{1}{2}\xi(1-\theta)}.$$

Plugging these in, we obtain

$$\mu^{+\hbar}=\frac{\frac{1}{2}\mu(1-\xi)(1-\theta)+\xi\mu\theta}{\xi\mu\theta+\frac{1}{2}(1-\xi)(1-\theta)}$$
$$\mu^{+\ell}=\frac{\frac{1}{2}\mu\xi(1-\theta)+(1-\xi)\mu\theta}{(1-\xi)\mu\theta+\frac{1}{2}\xi(1-\theta)}$$
$$\mu^{-\hbar}=\frac{\frac{1}{2}\mu(1-\xi)(1-\theta)}{\xi(1-\mu)\theta+\frac{1}{2}(1-\xi)(1-\theta)}$$
$$\mu^{-\ell}=\frac{\frac{1}{2}\mu\xi(1-\theta)}{(1-\xi)(1-\mu)\theta+\frac{1}{2}\xi(1-\theta)}$$

and

$$\mathbb{P}\{s=+,e=\hbar\}=\xi\mu\theta+\frac{1}{2}(1-\xi)(1-\theta)$$
$$\mathbb{P}\{s=+,e=\ell\}=(1-\xi)\mu\theta+\frac{1}{2}\xi(1-\theta)$$
$$\mathbb{P}\{s=-,e=\hbar\}=\xi(1-\mu)\theta+\frac{1}{2}(1-\xi)(1-\theta)$$
$$\mathbb{P}\{s=-,e=\ell\}=(1-\xi)(1-\mu)\theta+\frac{1}{2}\xi(1-\theta).$$

The first order derivative of the posteriors are

$$\frac{\partial}{\partial\xi}\mu^{+\hbar}=\frac{2\theta\mu(1-\theta)(1-\mu)}{(\theta\xi-\xi-\theta+2\theta\mu\xi+1)^2}>0$$
$$\frac{\partial}{\partial\xi}\mu^{+\ell}=\frac{-2\theta\mu(1-\theta)(1-\mu)}{(\xi+2\theta\mu-\theta\xi-2\theta\mu\xi)^2}<0$$
$$\frac{\partial}{\partial\xi}\mu^{-\hbar}=\frac{-2\theta\mu(1-\theta)(1-\mu)}{(\theta+\xi-3\theta\xi+2\theta\mu\xi-1)^2}<0$$
$$\frac{\partial}{\partial\xi}\mu^{-\ell}=\frac{2\theta\mu(1-\theta)(1-\mu)}{(2\theta+\xi-2\theta\mu-3\theta\xi+2\theta\mu\xi)^2}>0.$$

Therefore, $\mu^{+\hbar}$ and $\mu^{-\ell}$ are increasing while $\mu^{+\ell}$ and $\mu^{-\hbar}$ are decreasing in ξ. Furthermore, for $\xi=1/2$, $\mu^{+\hbar}=\mu^{+\ell}$ and $\mu^{-\hbar}=\mu^{-\ell}$ and for $\xi=1$, $\mu^{+\hbar}=1$, $\mu^{-\hbar}=0$ and $\mu^{+\ell}=\mu^{-\ell}=1/2$.

Proof of Proposition 2 By definition, complementarity requires

$$\mu^{+\hbar}_{|\xi=1/2}=\mu^{+}=\frac{\frac{1}{2}\mu\left(1-\theta\right)+\mu\theta}{\mu\theta+\frac{1}{2}\left(1-\theta\right)}<\overline{\mu}\Leftrightarrow\mu<\frac{\frac{1}{2}\overline{\mu}\left(1-\theta\right)}{\frac{1}{2}\left(1-\theta\right)+\underline{\mu}\theta}=\mu_H$$

and

$$\mu^{-\hbar}_{|\xi=1/2}=\mu^{-}=\frac{\frac{1}{2}\left(1-\theta\right)\mu}{\left(1-\mu\right)\theta+\frac{1}{2}\left(1-\theta\right)}>\underline{\mu}\Leftrightarrow\mu>\frac{\frac{1}{2}\underline{\mu}\left(1-\theta\right)+\underline{\mu}\theta}{\frac{1}{2}\left(1-\theta\right)+\underline{\mu}\theta}=\mu_L.$$

We define $I_\theta=(\mu_L,\mu_H)$. Note that $\mu_H+\mu_L=1$ for any $\lambda>0$ and $\theta\in(0,1)$. Note also that I_θ is not empty when $\mu_L<\mu_H$ or equivalently, when μ_L is lower than $\frac{1}{2}$ which happens when

$$\frac{\frac{1}{2}\underline{\mu}\left(1+\theta\right)}{\frac{1}{2}\left(1-\theta\right)+\theta\underline{\mu}}<\frac{1}{2}\Leftrightarrow\frac{1}{2}\underline{\mu}\left(1+\theta\right)<\frac{1}{4}\left(1-\theta\right)+\frac{1}{2}\theta\underline{\mu}$$
$$\Leftrightarrow\theta<\widehat{\theta}_\lambda=1-2\underline{\mu}=\frac{e^{1/\lambda}-1}{e^{1/\lambda}+1}.$$

Similarly, for a given $\mu\in(0,1)$ and $\theta\in(0,1)$,

$$\mu^{+}<\overline{\mu}\Leftrightarrow\mu^{+}<\frac{e^{1/\lambda}}{e^{1/\lambda}+1}\Leftrightarrow\lambda<\left(\ln\left(\frac{1}{1-\mu^{+}}-1\right)\right)^{-1}.$$

Similarly,

$$\mu^{-}>\underline{\mu}\Leftrightarrow\mu^{-}>\frac{1}{e^{1/\lambda}+1}\Leftrightarrow\lambda<\left(\ln\left(\frac{1}{\mu^{-}}-1\right)\right)^{-1}.$$

Therefore, the DM's judgment and the prediction of a black-box machine in the absence of explainability are complementary when

$$\lambda<\overline{\lambda}=\min\left\{\left(\ln\left(\frac{1}{1-\mu^{+}}-1\right)\right)^{-1},\left(\ln\left(\frac{1}{\mu^{-}}-1\right)\right)^{-1}\right\}.$$

Proof of Theorem 1 When $\xi=1/2$, we have $\underline{\mu}<\mu^{-\hbar}=\mu^{-\ell}<\mu<\mu^{+\ell}=\mu^{+\hbar}<\overline{\mu}$. Furthermore, per Lemma 1, as ξ increases, $\mu^{+\hbar}$ increases until 1, $\mu^{-\hbar}$ decreases until 0, and $\mu^{+\ell}$ decreases until μ and $\mu^{-\ell}$ increases until μ with ordering $\underline{\mu}<\mu^{-\hbar}<\mu^{-\ell}<\mu^{+\ell}<\mu^{+\hbar}<\overline{\mu}$ for $\xi>1/2$. Then this means that depending on the parameters, either $\mu^{+\hbar}$ will first hit the upper belief threshold $\overline{\mu}$ or $\mu^{-\hbar}$ will hit the lower belief threshold $\underline{\mu}$ first as ξ increases, where the former happens if and only if $\mu>0.5$. We label this point $\underline{\xi}$. Similarly, as ξ is further increased, the other posterior belief will hit the remaining threshold. We label this point $\overline{\xi}$. Note that the posteriors are symmetric when $\mu=0.5$, and hence $\underline{\xi}=\overline{\xi}$. For the remaining, we assume that $\mu>0.5$ and all results hold for $\mu<0.5$ since overreliance and underreliance are symmetric in μ around $\mu=0.5$.

Let $\xi<\underline{\xi}$. Then $\underline{\mu}<\mu^{-\hbar}<\mu^{-\ell}<\mu^{+\ell}<\mu^{+\hbar}<\overline{\mu}$. Note that overreliance given a positive machine prediction is

$$\begin{aligned}\mathbb{E}\left[\alpha_h^{\star}\left(\mu^{+E}\right)\mathbb{I}_{\{S=+\}}\right]&=\alpha_h^{\star}\left(\mu^{+\hbar}\right)\mathbb{P}\left\{s=+,e=\hbar\right\}+\alpha_h^{\star}\left(\mu^{+\ell}\right)\mathbb{P}\left\{s=+,e=\ell\right\}\\&=\mathbb{P}\left\{s=+\right\}\left(\alpha_h^{\star}\left(\mu^{+\hbar}\right)\mathbb{P}\left\{e=\hbar|s=+\right\}+\alpha_h^{\star}\left(\mu^{+\ell}\right)\mathbb{P}\left\{e=\ell|s=+\right\}\right)\\&=\mathbb{P}\left\{s=+\right\}\alpha_h^{\star}\left(\mu^{+}\right)\end{aligned}$$

where the latter is due to the linearity of $\alpha_h^{\star}$ on prior belief and

$$\mu^{+}=\mathbb{P}\left\{\omega=1|s=+\right\}=\mu^{+\hbar}\mathbb{P}\left\{e=\hbar|s=+\right\}+\mu^{+\ell}\mathbb{P}\left\{e=\ell|s=+\right\}.$$

Note that $\mathbb{P}\{s=+\}\,\alpha_h^\star(\mu^+)$ does not depend on explainability level ξ. In a similar fashion, overreliance given a negative machine output can be written as

$$\mathbb{E}\left[\beta_h^\star\left(\mu^{-E}\right)\mathbb{I}_{\{S=-\}}\right]=\mathbb{P}\{s=-\}\,\beta_h^\star(\mu^-)$$

which is also invariant to changes in ξ. Therefore overreliance, which is the sum of the two, is constant in ξ when $\xi<\underline{\xi}$.

Now assume that $\underline{\xi}\le\xi<\overline{\xi}$. Then we have $\underline{\mu}<\mu^{-\hbar}<\mu^{-\ell}<\mu^{+\ell}<\overline{\mu}$ and $\mu^{+\hbar}>\overline{\mu}$ since $\mu>0.5$. Now $\alpha_h^\star$ is no longer linear in prior belief. In particular, it is first increasing within $\left(\underline{\mu},\overline{\mu}\right)$ and then decreasing within the interval $[\overline{\mu},1]$. First, let us focus on $\alpha_h^\star\left(\mu^{+\hbar}\right)\mathbb{P}\{s=+,e=\hbar\}$. Note that

$$\begin{aligned}\alpha_h^\star\left(\mu^{+\hbar}\right)\mathbb{P}\{s=+,e=\hbar\}&=\left(1-\mu^{+\hbar}\right)\mathbb{P}\{s=+,e=\hbar\}\\&=\xi\mu\theta+\frac{1}{2}(1-\xi)(1-\theta)-\left(\frac{1}{2}\mu(1-\xi)(1-\theta)+\xi\mu\theta\right)\\&=\frac{1}{2}(1-\theta)(1-\mu)(1-\xi)\end{aligned}$$

which is decreasing in ξ. (Please see Proposition 1 for closed-form expressions for $\alpha_h^\star$ and $\beta_h^\star$.) Then this means that $\mathbb{E}[\alpha_h^\star(\mu^{+E})\mathbb{I}_{\{S=+\}}]$ is also decreasing since it is constant even when $\alpha_h^\star$ is evaluated when $\mu^{+\hbar}<\overline{\mu}$ where $\alpha_h^\star$ is increasing. In a similar manner, when $\xi>\overline{\xi}$, we have $\underline{\mu}<\mu^{-\ell}<\mu^{+\ell}<\overline{\mu}$ and $\mu^{+\hbar}>\overline{\mu}$ and $\mu^{-\hbar}<\underline{\mu}$. We focus on

$$\beta_h^\star\left(\mu^{-\hbar}\right)\mathbb{P}\{s=-,e=\hbar\}=\beta_h^\star\left(\mu^{-\hbar}\right)\mathbb{P}\{s=-,e=\hbar\}=\mu^{-\hbar}\mathbb{P}\{s=-,e=\hbar\}=\frac{1}{2}\mu(1-\xi)(1-\theta)$$

which is decreasing in ξ. Note again that $\mathbb{E}[\beta_h^\star(\mu^{-E})\mathbb{I}_{\{S=-\}}]$ is constant even when $\beta_h^\star$ is evaluated at $\mu^{-\hbar}\in(\underline{\mu},\overline{\mu})$ for $\xi<\overline{\xi}$ which was decreasing in μ and hence increasing in ξ since $\mu^{+\ell}$ is decreasing in ξ. Since $\beta_h^\star\left(\mu^{-\hbar}\right)\mathbb{P}\{s=-,e=\hbar\}$ is decreasing for $\xi>\overline{\xi}$, it follows that $\mathbb{E}[\beta_h^\star(\mu^{-E})\mathbb{I}_{\{S=-\}}]$ is also decreasing. This also means that the rate of decrease is higher for $\xi>\overline{\xi}$ compared to when $\underline{\xi}\le\xi<\overline{\xi}$.

The same arguments apply to underreliance. However, since false positives and false negatives are reversed for $U(\xi)$, the impact of ξ is also reversed and $U(\xi)$ is increasing in ξ. In a similar fashion, appropriate reliance, or expected accuracy, is constant in the first interval $\xi<\underline{\xi}$ since $A_h^\star$ is constant within $[\underline{\mu},\overline{\mu}]$. More specifically, it is equal to

$$A(\xi)\equiv\mathbb{E}\left[A_h^\star\left(\mu^{SE}\right)\right]=\frac{e^{1/\lambda}}{e^{1/\lambda}+1}.$$

Now, assume that $\xi\ge\underline{\xi}$. Then $\mu^{+\hbar}\ge\overline{\mu}$. Note that $A_h^\star(\mu)$ is increasing in μ within $[\overline{\mu},1]$ (see Proposition 1). Then note that we can write expected accuracy as

$$\begin{aligned}A(\xi)&=\frac{e^{1/\lambda}}{e^{1/\lambda}+1}+\mathbb{P}\{s=+,e=\hbar\}\left(A_h^\star\left(\mu^{+\hbar}\right)-\frac{e^{1/\lambda}}{e^{1/\lambda}+1}\right)\\&=\frac{e^{1/\lambda}}{e^{1/\lambda}+1}+\left(\frac{1}{2}\mu(1-\xi)(1-\theta)+\xi\mu\theta-\left(\xi\mu\theta+\frac{1}{2}(1-\xi)(1-\theta)\right)\frac{e^{1/\lambda}}{e^{1/\lambda}+1}\right)\end{aligned}$$

Its first order derivative is

$$-\frac{1}{2}\mu(1-\theta)+\mu\theta-\left(\mu\theta-\frac{1}{2}(1-\theta)\right)\frac{e^{1/\lambda}}{e^{1/\lambda}+1}$$

which is linear in θ. For $\theta=0$, it is equal to

$$-\frac{1}{2}\mu+\frac{1}{2}\frac{e^{1/\lambda}}{e^{1/\lambda}+1}>0$$

since $\mu < \overline{\mu} = \frac{e^{1/\lambda}}{e^{1/\lambda}+1}$ by Lemma 2. For $\theta = 1$, it is equal to

$$\mu \frac{1}{e^{1/\lambda}+1} > 0.$$

Since the first order derivative is positive at the extremes, it is always positive due to linearity. Therefore, accuracy is increasing.

Assume now that $\xi \geq \overline{\xi}$. Then $\mu^{-\hbar} < \underline{\mu}$. Accuracy then can be written as

$$\begin{aligned} A(\xi) &= \frac{e^{1/\lambda}}{e^{1/\lambda}+1} + \mathbb{P}\{s=+, e=\hbar\}\left(A_h^\star\left(\mu^{+\hbar}\right) - \frac{e^{1/\lambda}}{e^{1/\lambda}+1}\right) + \mathbb{P}\{s=-, e=\hbar\}\left(A_h^\star\left(\mu^{-\hbar}\right) - \frac{e^{1/\lambda}}{e^{1/\lambda}+1}\right) \\ &= \frac{e^{1/\lambda}}{e^{1/\lambda}+1} + \mathbb{P}\{s=+, e=\hbar\}\left(\mu^{+\hbar} - \frac{e^{1/\lambda}}{e^{1/\lambda}+1}\right) + \mathbb{P}\{s=-, e=\hbar\}\left(1 - \mu^{-\hbar} - \frac{e^{1/\lambda}}{e^{1/\lambda}+1}\right) \end{aligned}$$

We have already shown that the second component is decreasing in ξ. We focus on the third component. It is equal to

$$\begin{aligned} &\mathbb{P}\{s=-, e=\hbar\}\left(1 - \mu^{-\hbar} - \frac{e^{1/\lambda}}{e^{1/\lambda}+1}\right) \\ &= \left(\xi(1-\mu)\theta + \frac{1}{2}(1-\xi)(1-\theta)\right) - \left(\frac{1}{2}\mu(1-\xi)(1-\theta)\right) - \left(\xi(1-\mu)\theta + \frac{1}{2}(1-\xi)(1-\theta)\right)\frac{e^{1/\lambda}}{e^{1/\lambda}+1}. \end{aligned}$$

Its first-order derivative is

$$\left((1-\mu)\theta - \frac{1}{2}(1-\theta)\right) + \frac{1}{2}\mu(1-\theta) - \left((1-\mu)\theta - \frac{1}{2}(1-\theta)\right)\frac{e^{1/\lambda}}{e^{1/\lambda}+1}$$

which is linear in θ. For $\theta = 0$, it is equal to

$$-\frac{1}{2}\frac{1}{e^{1/\lambda}+1} + \frac{1}{2}\mu$$

which is always positive since $\mu > \underline{\mu} = \frac{1}{e^{1/\lambda}+1}$ by Lemma 2. For $\theta = 1$, it is equal to

$$(1-\mu)\frac{1}{e^{1/\lambda}+1}$$

which is always positive. Since the first order derivative is positive at the extremes, it is always positive due to linearity. Therefore, underreliance is increasing in ξ with a higher rate of increase when $\xi \geq \overline{\xi}$.

Proof of Corollary 1 Assume that $\xi = 1$. In this case, $\mu^{+\hbar} = 1$, $\mu^{-\hbar} = 0$ and $\mu^{+\ell} = \mu^{-\ell} = \mu$ and

$$\begin{aligned} \mathbb{P}\{s=+, e=\hbar\} &= \mu\theta \\ \mathbb{P}\{s=+, e=\ell\} &= \frac{1}{2}(1-\theta) \\ \mathbb{P}\{s=-, e=\hbar\} &= (1-\mu)\theta \\ \mathbb{P}\{s=-, e=\ell\} &= \frac{1}{2}(1-\theta). \end{aligned}$$

Then overreliance is

$$O(\xi) = \left(1-\mu^{+\hbar}\right)\mathbb{P}\{s=+, e=\hbar\} + \frac{\mu\left(e^{1/\lambda}+1\right)-1}{e^{2/\lambda}-1}\mathbb{P}\{s=+, e=\ell\} + \frac{e^{1/\lambda} - \mu\left(e^{1/\lambda}+1\right)}{e^{2/\lambda}-1}\mathbb{P}\{s=-, e=\ell\}$$

which simplifies to

$$O(\xi) = \frac{1-\theta}{2e^{\frac{1}{\lambda}}+2}.$$

Similarly, underreliance is

$$\begin{aligned} U(\xi) &= \beta_h^\star\left(\mu^{+\ell}\right)\mathbb{P}\{s=+, e=\ell\} + \alpha_h^\star\left(\mu^{-\ell}\right)\mathbb{P}\{s=-, e=\ell\} \\ &= \frac{e^{1/\lambda} - \mu\left(e^{1/\lambda}+1\right)}{e^{2/\lambda}-1}\frac{1}{2}(1-\theta) + \frac{\mu\left(e^{1/\lambda}+1\right)-1}{e^{2/\lambda}-1}\frac{1}{2}(1-\theta) = \frac{1-\theta}{2e^{\frac{1}{\lambda}}+2}. \end{aligned}$$

That is, overreliance and underreliance are equal to each other when $\xi = 1$. Since, per Theorem 1, overreliance is decreasing and underreliance is increasing in ξ, overreliance is always higher than underreliance.

Proof of Proposition 3 Assume that $\mu > 1/2$. Per Theorem 1, $\Delta O \leq 0$, $\Delta U \geq 0$ and $\Delta A \geq 0$. More specifically, overreliance, underreliance, and accuracy are constant in ξ, and hence $\Delta O = 0$, $\Delta U = 0$, and $\Delta A = 0$ for $\xi \leq \underline{\xi}$. Note also that

$$\xi \leq \underline{\xi} \Leftrightarrow \mu^{+\hbar} < \overline{\mu} = \frac{e^{1/\lambda}}{e^{1/\lambda}+1} \Leftrightarrow \lambda \leq 1/\ln \frac{\mu^{+\hbar}}{1-\mu^{+\hbar}} = \widehat{\lambda}.$$

Assume that $\underline{\xi} < \xi < \overline{\xi}$. This means that $\widehat{\lambda} < \lambda < \widehat{\lambda}_1$ for some $\widehat{\lambda}_1 > \widehat{\lambda}$. Then $\mu^{+\hbar} > \overline{\mu}$ and $\underline{\mu} < \mu^{-\hbar} < \mu^{-\ell} < \mu^{+\ell} < \overline{\mu}$. Note that ΔO in this case is simply

$$\Delta O = \left(\left(1-\mu^{+\hbar}\right) - \frac{\mu^{+\hbar}\left(e^{1/\lambda}+1\right)-1}{e^{2/\lambda}-1} \right) \mathbb{P}\{s=+, e=\hbar\}$$

where $1-\mu^{+\hbar}$ is the false-positive error when $\mu^{+\hbar} > \overline{\mu}$ and the second term inside parentheses is the false-positive error when the DM puts effort, i.e., $\underline{\mu} < \mu^{+\hbar} < \overline{\mu}$, since $\Delta O = 0$ when the DM puts effort for all posteriors. Plugging in the terms, we have

$$\begin{aligned}\Delta O = &\left(\xi\mu\theta + \frac{1}{2}(1-\xi)(1-\theta) - \left(\frac{1}{2}\mu(1-\xi)(1-\theta)+\xi\mu\theta\right)\right) \\ &- \frac{\left(\frac{1}{2}\mu(1-\xi)(1-\theta)+\xi\mu\theta\right)\left(e^{1/\lambda}+1\right) - \left(\xi\mu\theta+\frac{1}{2}(1-\xi)(1-\theta)\right)}{e^{2/\lambda}-1}.\end{aligned}$$

Its first order derivative with respect to λ is

$$\begin{aligned}\frac{\partial}{\partial\lambda}\Delta O &= \frac{\partial}{\partial\lambda}\left(\begin{array}{c} \left(\xi\mu\theta+\frac{1}{2}(1-\xi)(1-\theta)-\left(\frac{1}{2}\mu(1-\xi)(1-\theta)+\xi\mu\theta\right)\right) \\ -\frac{\left(\frac{1}{2}\mu(1-\xi)(1-\theta)+\xi\mu\theta\right)\left(e^{1/\lambda}+1\right)-\left(\xi\mu\theta+\frac{1}{2}(1-\xi)(1-\theta)\right)}{e^{2/\lambda}-1} \end{array} \right) \\ &= \frac{\left(\begin{array}{c} 2e^{2\frac{1}{\lambda}} - 2\mu e^{2\frac{1}{\lambda}} - \mu e^{3\frac{1}{\lambda}} - 2\xi e^{2\frac{1}{\lambda}} - 2\theta e^{2\frac{1}{\lambda}} - \mu e^{\frac{1}{\lambda}} + \theta\mu e^{\frac{1}{\lambda}} + \mu\xi e^{\frac{1}{\lambda}} \\ +2\theta\mu e^{2\frac{1}{\lambda}} + \theta\mu e^{3\frac{1}{\lambda}} + 2\theta\xi e^{2\frac{1}{\lambda}} + 2\mu\xi e^{2\frac{1}{\lambda}} + \mu\xi e^{3\frac{1}{\lambda}} - 3\theta\mu\xi e^{\frac{1}{\lambda}} - 2\theta\mu\xi e^{2\frac{1}{\lambda}} - 3\theta\mu\xi e^{3\frac{1}{\lambda}} \end{array} \right)}{2\lambda^2\left(e^{2/\lambda}-1\right)^2}.\end{aligned}$$

Note that the denominator is always positive. We focus on the numerator. It is linear in θ. For $\theta = 0$, we have

$$2e^{2\frac{1}{\lambda}} - 2\mu e^{2\frac{1}{\lambda}} - \mu e^{3\frac{1}{\lambda}} - 2\xi e^{2\frac{1}{\lambda}} - \mu e^{\frac{1}{\lambda}} + \mu\xi e^{\frac{1}{\lambda}} + 2\mu\xi e^{2\frac{1}{\lambda}} + \mu\xi e^{3\frac{1}{\lambda}} = -e^{\frac{1}{\lambda}}(1-\xi)\left(\mu + \mu e^{2\frac{1}{\lambda}} + 2\mu e^{\frac{1}{\lambda}} - 2e^{\frac{1}{\lambda}}\right)$$

which is also linear in μ. For $\mu = \underline{\mu} = \frac{1}{e^{1/\lambda}+1}$, it becomes

$$-e^{\frac{1}{\lambda}}(1-\xi)\left(\frac{1}{e^{1/\lambda}+1} + \frac{1}{e^{1/\lambda}+1}e^{2\frac{1}{\lambda}} + 2\frac{1}{e^{1/\lambda}+1}e^{\frac{1}{\lambda}} - 2e^{\frac{1}{\lambda}}\right) = -e^{\frac{1}{\lambda}}(1-\xi)\left(1-e^{\frac{1}{\lambda}}\right) < 0$$

which is negative. For $\mu = \overline{\mu}$, it becomes

$$-e^{\frac{1}{\lambda}}(1-\xi)\left(\frac{e^{1/\lambda}}{e^{1/\lambda}+1} + \frac{e^{1/\lambda}}{e^{1/\lambda}+1}e^{2\frac{1}{\lambda}} + 2\frac{e^{1/\lambda}}{e^{1/\lambda}+1}e^{\frac{1}{\lambda}} - 2e^{\frac{1}{\lambda}}\right) = -e^{\frac{1}{\lambda}}(1-\xi)e^{\frac{1}{\lambda}}\left(e^{\frac{1}{\lambda}}-1\right) < 0$$

which is also negative. Therefore the numerator is negative for $\theta = 0$ due to linearity in μ. Similarly assume that $\theta = 1$. Then the numerator becomes

$$\begin{aligned}&2e^{2\frac{1}{\lambda}} - 2\mu e^{2\frac{1}{\lambda}} - \mu e^{3\frac{1}{\lambda}} - 2\xi e^{2\frac{1}{\lambda}} - 2e^{2\frac{1}{\lambda}} - \mu e^{\frac{1}{\lambda}} + \mu e^{\frac{1}{\lambda}} + \mu\xi e^{\frac{1}{\lambda}} \\ &+ 2\mu e^{2\frac{1}{\lambda}} + \mu e^{3\frac{1}{\lambda}} + 2\xi e^{2\frac{1}{\lambda}} + 2\mu\xi e^{2\frac{1}{\lambda}} + \mu\xi e^{3\frac{1}{\lambda}} - 3\mu\xi e^{\frac{1}{\lambda}} - 2\mu\xi e^{2\frac{1}{\lambda}} - 3\mu\xi e^{3\frac{1}{\lambda}} \\ &= -2\mu\xi e^{\frac{1}{\lambda}}\left(e^{2\frac{1}{\lambda}}+1\right) < 0\end{aligned}$$

after simplification which is also negative. Since the numerator is negative for extreme values of θ, it is negative for all values due to linearity in θ. Thus, $\frac{\partial}{\partial\lambda}\Delta O<0$.

In a similar fashion, the change in underreliance is

$$\Delta U=-\frac{e^{1/\lambda}-\mu^{+\hbar}\left(e^{1/\lambda}+1\right)}{e^{2/\lambda}-1}\mathbb{P}\{s=+,e=\hbar\}$$

since $\beta_h^\star\left(\mu^{+\hbar}\right)=0$ when $\mu^{+\hbar}>\overline{\mu}$. Plugging in terms, we obtain

$$\Delta U=-\frac{\left(\xi\mu\theta+\frac{1}{2}\left(1-\xi\right)\left(1-\theta\right)\right)-\left(\frac{1}{2}\mu\left(1-\xi\right)\left(1-\theta\right)+\xi\mu\theta\right)\left(e^{1/\lambda}+1\right)}{e^{2/\lambda}-1}$$

whose first order derivative with respect to λ is

$$\begin{aligned}\frac{\partial}{\partial\lambda}\Delta U&=-\frac{\partial}{\partial\lambda}\frac{\left(\xi\mu\theta+\frac{1}{2}\left(1-\xi\right)\left(1-\theta\right)\right)-\left(\frac{1}{2}\mu\left(1-\xi\right)\left(1-\theta\right)+\xi\mu\theta\right)\left(e^{1/\lambda}+1\right)}{e^{2/\lambda}-1}\\&=-\frac{\begin{pmatrix}2e^{2\frac{1}{\lambda}}-2\mu e^{2\frac{1}{\lambda}}-\mu e^{3\frac{1}{\lambda}}-2\xi e^{2\frac{1}{\lambda}}-2\theta e^{2\frac{1}{\lambda}}-\mu e^{\frac{1}{\lambda}}+\theta\mu e^{\frac{1}{\lambda}}+\mu\xi e^{\frac{1}{\lambda}}+2\theta\mu e^{2\frac{1}{\lambda}}\\+\theta\mu e^{3\frac{1}{\lambda}}+2\theta\xi e^{2\frac{1}{\lambda}}+2\mu\xi e^{2\frac{1}{\lambda}}+\mu\xi e^{3\frac{1}{\lambda}}-3\theta\mu\xi e^{\frac{1}{\lambda}}-2\theta\mu\xi e^{2\frac{1}{\lambda}}-3\theta\mu\xi e^{3\frac{1}{\lambda}}\end{pmatrix}}{2\lambda^2\left(e^{2/\lambda}-1\right)^2}.\end{aligned}$$

Since the denominator is positive, we focus on the numerator which is linear in θ. For $\theta=0$, it becomes

$$2e^{2\frac{1}{\lambda}}-2\mu e^{2\frac{1}{\lambda}}-\mu e^{3\frac{1}{\lambda}}-2\xi e^{2\frac{1}{\lambda}}-\mu e^{\frac{1}{\lambda}}+\mu\xi e^{\frac{1}{\lambda}}+2\mu\xi e^{2\frac{1}{\lambda}}+\mu\xi e^{3\frac{1}{\lambda}}=-e^{\frac{1}{\lambda}}\left(1-\xi\right)\left(\mu-2e^{\frac{1}{\lambda}}+\mu e^{2\frac{1}{\lambda}}+2\mu e^{\frac{1}{\lambda}}\right)$$

which is linear in μ. For $\mu=\underline{\mu}$, it becomes $-e^{\frac{1}{\lambda}}\left(1-\xi\right)\left(1-e^{\frac{1}{\lambda}}\right)<0$ which is negative. For $\mu=\overline{\mu}$, it becomes $-e^{\frac{1}{\lambda}}\left(1-\xi\right)e^{\frac{1}{\lambda}}\left(e^{\frac{1}{\lambda}}-1\right)<0$ which is also negative. Since it is negative for both endpoints for μ, it is always positive for $\theta=0$ due to its linearity in θ. For $\theta=1$, the numerator becomes

$$\begin{aligned}&2e^{2\frac{1}{\lambda}}-2\mu e^{2\frac{1}{\lambda}}-\mu e^{3\frac{1}{\lambda}}-2\xi e^{2\frac{1}{\lambda}}-2e^{2\frac{1}{\lambda}}-\mu e^{\frac{1}{\lambda}}+\mu e^{\frac{1}{\lambda}}+\mu\xi e^{\frac{1}{\lambda}}+2\mu e^{2\frac{1}{\lambda}}\\&+\mu e^{3\frac{1}{\lambda}}+2\xi e^{2\frac{1}{\lambda}}+2\mu\xi e^{2\frac{1}{\lambda}}+\mu\xi e^{3\frac{1}{\lambda}}-3\mu\xi e^{\frac{1}{\lambda}}-2\mu\xi e^{2\frac{1}{\lambda}}-3\mu\xi e^{3\frac{1}{\lambda}}=-2\mu\xi e^{\frac{1}{\lambda}}\left(e^{2\frac{1}{\lambda}}+1\right)<0\end{aligned}$$

which is always negative. Since it is negative for both endpoints for θ, it is always negative due to its linearity in θ. Therefore $\frac{\partial}{\partial\lambda}\Delta U>0$ and hence underreliance is increasing in λ.

Finally, the change in accuracy can be written as

$$\Delta A=\left(\mu^{+\hbar}-\frac{e^{1/\lambda}}{e^{1/\lambda}+1}\right)\mathbb{P}\{s=+,e=\hbar\}=\left(\frac{1}{2}\mu\left(1-\xi\right)\left(1-\theta\right)+\xi\mu\theta\right)-\frac{e^{1/\lambda}}{e^{1/\lambda}+1}\left(\xi\mu\theta+\frac{1}{2}\left(1-\xi\right)\left(1-\theta\right)\right)$$

which is increasing in λ since $\frac{e^{1/\lambda}}{e^{1/\lambda}+1}=1-\frac{1}{e^{1/\lambda}+1}$ is decreasing in λ.

Assume now that $\xi>\overline{\xi}$. This means that $\lambda\geq\widehat{\lambda}_1$. Then $\mu^{+\hbar}>\overline{\mu}$, $\mu^{-\hbar}<\underline{\mu}$ and $\underline{\mu}<\mu^{-\ell}<\mu^{+\ell}<\overline{\mu}$. Then, ΔO in this case is simply

$$\Delta O=\left(\left(1-\mu^{+\hbar}\right)-\frac{\mu^{+\hbar}\left(e^{1/\lambda}+1\right)-1}{e^{2/\lambda}-1}\right)\mathbb{P}\{s=+,e=\hbar\}+\left(\mu^{-\hbar}-\frac{e^{1/\lambda}-\mu^{-\hbar}\left(e^{1/\lambda}+1\right)}{e^{2/\lambda}-1}\right)\mathbb{P}\{s=-,e=\hbar\}. \tag{1}$$

We already know that the first term is decreasing in λ. Let us focus on the second term, which becomes

$$\begin{aligned}&\left(\mu^{-\hbar}-\frac{e^{1/\lambda}-\mu^{-\hbar}\left(e^{1/\lambda}+1\right)}{e^{2/\lambda}-1}\right)\mathbb{P}\left\{s=-,e=\hbar\right\}\\&=\left(\frac{1}{2}\mu\left(1-\xi\right)\left(1-\theta\right)\right)-\frac{e^{1/\lambda}\left(\xi\left(1-\mu\right)\theta+\frac{1}{2}\left(1-\xi\right)\left(1-\theta\right)\right)-\left(\frac{1}{2}\mu\left(1-\xi\right)\left(1-\theta\right)\right)\left(e^{1/\lambda}+1\right)}{e^{2/\lambda}-1}\end{aligned}$$

after plugging in the terms. Its first order derivative with respect to λ is

$$\frac{\partial}{\partial\lambda}\left(-\frac{e^{1/\lambda}\left(\xi(1-\mu)\theta+\frac{1}{2}(1-\xi)(1-\theta)\right)-\left(\frac{1}{2}\mu(1-\xi)(1-\theta)\right)\left(e^{1/\lambda}+1\right)}{e^{2/\lambda}-1}\right)$$
$$-\frac{\left(\begin{array}{c} e^{\frac{1}{\lambda}}-\theta e^{3\frac{1}{\lambda}}-2\mu e^{2\frac{1}{\lambda}}-\mu e^{3\frac{1}{\lambda}}-\xi e^{3\frac{1}{\lambda}}+e^{3\frac{1}{\lambda}}-\theta e^{\frac{1}{\lambda}}-\mu e^{\frac{1}{\lambda}}-\xi e^{\frac{1}{\lambda}}+\theta\mu e^{\frac{1}{\lambda}}+3\theta\xi e^{\frac{1}{\lambda}} \\ +\mu\xi e^{\frac{1}{\lambda}}+2\theta\mu e^{2\frac{1}{\lambda}}+\theta\mu e^{3\frac{1}{\lambda}}+3\theta\xi e^{3\frac{1}{\lambda}}+2\mu\xi e^{2\frac{1}{\lambda}}+\mu\xi e^{3\frac{1}{\lambda}}-3\theta\mu\xi e^{\frac{1}{\lambda}}-2\theta\mu\xi e^{2\frac{1}{\lambda}}-3\theta\mu\xi e^{3\frac{1}{\lambda}}\end{array}\right)}{2\lambda^2\left(e^{2/\lambda}-1\right)^2}. \tag{2}$$

Since the denominator is positive, we again focus on the numerator, which is linear in θ. For $\theta=0$, it becomes

$$\begin{aligned}&e^{\frac{1}{\lambda}}-2\mu e^{2\frac{1}{\lambda}}-\mu e^{3\frac{1}{\lambda}}-\xi e^{3\frac{1}{\lambda}}+e^{3\frac{1}{\lambda}}-\mu e^{\frac{1}{\lambda}}-\xi e^{\frac{1}{\lambda}}+\mu\xi e^{\frac{1}{\lambda}}+2\mu\xi e^{2\frac{1}{\lambda}}+\mu\xi e^{3\frac{1}{\lambda}}\\ &=-e^{\frac{1}{\lambda}}(1-\xi)\left(\mu+\mu e^{2\frac{1}{\lambda}}-e^{2\frac{1}{\lambda}}+2\mu e^{\frac{1}{\lambda}}-1\right)\end{aligned} \tag{3}$$

which is linear in ξ. For $\xi=1$, it is zero. Let us evaluate it for the other extreme where $\xi=\overline{\xi}$. Note that $\overline{\xi}$ can be found by solving $\mu^{-\hbar}=\underline{\mu}$ which is equivalent to $\mu=\underline{\mu}$ for $\theta=0$. At $\mu=\underline{\mu}$, (3) becomes

$$-e^{\frac{1}{\lambda}}(1-\xi)\left(\frac{1}{e^{\frac{1}{\lambda}}+1}+\frac{1}{e^{\frac{1}{\lambda}}+1}e^{2\frac{1}{\lambda}}-e^{2\frac{1}{\lambda}}+2\frac{1}{e^{\frac{1}{\lambda}}+1}e^{\frac{1}{\lambda}}-1\right)=e^{\frac{1}{\lambda}}\left(1-e^{\frac{1}{\lambda}}\right)>0$$

which is positive. Therefore, it is always positive for $\theta=0$ due to linearity in ξ. Similarly, for $\theta=1$, the numerator in (2) becomes

$$\begin{aligned}&e^{\frac{1}{\lambda}}-e^{3\frac{1}{\lambda}}-2\mu e^{2\frac{1}{\lambda}}-\mu e^{3\frac{1}{\lambda}}-\xi e^{3\frac{1}{\lambda}}+e^{3\frac{1}{\lambda}}-e^{\frac{1}{\lambda}}-\mu e^{\frac{1}{\lambda}}-\xi e^{\frac{1}{\lambda}}+\mu e^{\frac{1}{\lambda}}+3\xi e^{\frac{1}{\lambda}}+\mu\xi e^{\frac{1}{\lambda}}\\ &+2\mu e^{2\frac{1}{\lambda}}+\mu e^{3\frac{1}{\lambda}}+3\xi e^{3\frac{1}{\lambda}}+2\mu\xi e^{2\frac{1}{\lambda}}+\mu\xi e^{3\frac{1}{\lambda}}-3\mu\xi e^{\frac{1}{\lambda}}-2\mu\xi e^{2\frac{1}{\lambda}}-3\mu\xi e^{3\frac{1}{\lambda}}=2\xi e^{\frac{1}{\lambda}}\left(e^{2\frac{1}{\lambda}}+1\right)(1-\mu)>0\end{aligned}$$

which is always positive. Since the numerator in (2) is positive at extreme values of θ, it is always positive since it is linear in θ. Since the denominator is positive and there is a negative sign at the beginning of (2), the first order derivative of the second term in (1) is negative. Since we already know that the first term is also negative, the first order derivative is negative and hence ΔO is decreasing in λ where the rate of increase is higher for $\lambda>\widehat{\lambda}_1$.

By the same token, the change in underreliance is

$$-\frac{e^{1/\lambda}-\mu^{+\hbar}\left(e^{1/\lambda}+1\right)}{e^{2/\lambda}-1}\mathbb{P}\{s=+,e=\hbar\}-\frac{\mu^{-\hbar}\left(e^{1/\lambda}+1\right)-1}{e^{2/\lambda}-1}\mathbb{P}\{s=-,e=\hbar\} \tag{4}$$

since $\beta_{\hbar}^{\star}\left(\mu^{+\hbar}\right)=0$ for $\mu^{+\hbar}>\overline{\mu}$ and $\alpha_{\hbar}^{\star}\left(\mu^{-\hbar}\right)=0$ for $\mu^{-\hbar}<\underline{\mu}$. We already know that the first term is increasing in λ. Hence we focus on the second term which, after plugging in $\mu^{-\hbar}$, becomes

$$-\frac{\mu^{-\hbar}\left(e^{1/\lambda}+1\right)-1}{e^{2/\lambda}-1}\mathbb{P}\{s=-,e=\hbar\}=-\frac{\frac{1}{2}\mu(1-\xi)(1-\theta)\left(e^{1/\lambda}+1\right)-\left(\xi(1-\mu)\theta+\frac{1}{2}(1-\xi)(1-\theta)\right)}{e^{2/\lambda}-1}$$

whose first order derivative with respect to λ is

$$\begin{aligned}&\frac{\partial}{\partial\lambda}\left(-\frac{\frac{1}{2}\mu(1-\xi)(1-\theta)\left(e^{1/\lambda}+1\right)-\left(\xi(1-\mu)\theta+\frac{1}{2}(1-\xi)(1-\theta)\right)}{e^{2/\lambda}-1}\right)\\ &=\frac{\left(\begin{array}{c}2e^{2\frac{1}{\lambda}}-2\mu e^{2\frac{1}{\lambda}}-\mu e^{3\frac{1}{\lambda}}-2\xi e^{2\frac{1}{\lambda}}-2\theta e^{2\frac{1}{\lambda}}-\mu e^{\frac{1}{\lambda}}+\theta\mu e^{\frac{1}{\lambda}}+\mu\xi e^{\frac{1}{\lambda}}\\ +2\theta\mu e^{2\frac{1}{\lambda}}+\theta\mu e^{3\frac{1}{\lambda}}+6\theta\xi e^{2\frac{1}{\lambda}}+2\mu\xi e^{2\frac{1}{\lambda}}+\mu\xi e^{3\frac{1}{\lambda}}-\theta\mu\xi e^{\frac{1}{\lambda}}-6\theta\mu\xi e^{2\frac{1}{\lambda}}-\theta\mu\xi e^{3\frac{1}{\lambda}}\end{array}\right)}{2\lambda^2\left(e^{2/\lambda}-1\right)^2}.\end{aligned}$$

Since the denominator is positive, we focus on the numerator, which is linear in θ. For $\theta=0$ it becomes

$$\left(2e^{2\frac{1}{\lambda}}-2\mu e^{2\frac{1}{\lambda}}-\mu e^{3\frac{1}{\lambda}}-2\xi e^{2\frac{1}{\lambda}}-\mu e^{\frac{1}{\lambda}}+\mu\xi e^{\frac{1}{\lambda}}+2\mu\xi e^{2\frac{1}{\lambda}}+\mu\xi e^{3\frac{1}{\lambda}}\right)=e^{\frac{1}{\lambda}}(1-\xi)\left(\mu-2e^{\frac{1}{\lambda}}+\mu e^{2\frac{1}{\lambda}}+2\mu e^{\frac{1}{\lambda}}\right)$$

which is linear in μ. For $\mu = \underline{\mu}$ it becomes

$$e^{\frac{1}{\lambda}}(1-\xi)\left(\frac{1}{e^{\frac{1}{\lambda}}+1} - 2e^{\frac{1}{\lambda}} + \frac{1}{e^{\frac{1}{\lambda}}+1}e^{2\frac{1}{\lambda}} + 2\frac{1}{e^{\frac{1}{\lambda}}+1}e^{\frac{1}{\lambda}}\right) = e^{\frac{1}{\lambda}}(1-\xi)\left(1-e^{\frac{1}{\lambda}}\right) > 0.$$

which is positive. For $\mu = \overline{\mu}$, it becomes

$$e^{\frac{1}{\lambda}}(1-\xi)\left(\frac{e^{\frac{1}{\lambda}}}{e^{\frac{1}{\lambda}}+1} - 2e^{\frac{1}{\lambda}} + \frac{e^{\frac{1}{\lambda}}}{e^{\frac{1}{\lambda}}+1}e^{2\frac{1}{\lambda}} + 2\frac{e^{\frac{1}{\lambda}}}{e^{\frac{1}{\lambda}}+1}e^{\frac{1}{\lambda}}\right) = e^{2\frac{1}{\lambda}}(1-\xi)\left(e^{\frac{1}{\lambda}}-1\right) > 0$$

which is also positive. Since it is positive at both endpoints for μ, it is positive for $\theta = 0$ due to its linearity in θ. Assume that $\theta = 1$. Then the numerator becomes

$$\begin{gathered} 2e^{2\frac{1}{\lambda}} - 2\mu e^{2\frac{1}{\lambda}} - \mu e^{3\frac{1}{\lambda}} - 2\xi e^{2\frac{1}{\lambda}} - 2e^{2\frac{1}{\lambda}} - \mu e^{\frac{1}{\lambda}} + \mu e^{\frac{1}{\lambda}} + \mu\xi e^{\frac{1}{\lambda}} + 2\mu e^{2\frac{1}{\lambda}} \\ +\mu e^{3\frac{1}{\lambda}} + 6\xi e^{2\frac{1}{\lambda}} + 2\mu\xi e^{2\frac{1}{\lambda}} + \mu\xi e^{3\frac{1}{\lambda}} - \mu\xi e^{\frac{1}{\lambda}} - 6\mu\xi e^{2\frac{1}{\lambda}} - \mu\xi e^{3\frac{1}{\lambda}} = 4\xi e^{\frac{2}{\lambda}}(1-\mu) > 0 \end{gathered}$$

which is always positive. Therefore, the numerator is always positive due to linearity in θ, and it is positive at both endpoints for θ. Therefore, the second term in (4) is also positive. This means that ΔU is increasing in λ with a higher rate of increase for $\lambda > \widehat{\lambda}_1$ than $\widehat{\lambda} < \lambda < \widehat{\lambda}_1$.

Finally, we focus on the expected accuracy. The change in accuracy is

$$\Delta A = \left(\mu^{+\hbar} - \frac{e^{1/\lambda}}{e^{1/\lambda}+1}\right)\mathbb{P}\{s=+, e=\hbar\} + \left(1 - \mu^{-\hbar} - \frac{e^{1/\lambda}}{e^{1/\lambda}+1}\right)\mathbb{P}\{s=-, e=\hbar\}.$$

Again, we already know that the first term is increasing in λ. Let us focus on the second term. It becomes

$$\left(\xi(1-\mu)\theta + \frac{1}{2}(1-\xi)(1-\theta)\right) - \left(\frac{1}{2}\mu(1-\xi)(1-\theta)\right) - \frac{e^{1/\lambda}}{e^{1/\lambda}+1}\left(\xi(1-\mu)\theta + \frac{1}{2}(1-\xi)(1-\theta)\right)$$

whose first order derivative with respect to λ is

$$\begin{aligned} &\frac{\partial}{\partial\lambda}\left(\left(\xi(1-\mu)\theta + \frac{1}{2}(1-\xi)(1-\theta)\right) - \left(\frac{1}{2}\mu(1-\xi)(1-\theta)\right) - \frac{e^{1/\lambda}}{e^{1/\lambda}+1}\left(\xi(1-\mu)\theta + \frac{1}{2}(1-\xi)(1-\theta)\right)\right) \\ &-\frac{e^{\frac{1}{\lambda}}(\theta + \xi - 3\theta\xi + 2\theta\mu\xi - 1)}{2\lambda^2 e^{2\frac{1}{\lambda}} + 4\lambda^2 e^{\frac{1}{\lambda}} + 2\lambda^2} \end{aligned} \tag{5}$$

Since the denominator is positive, we focus on the numerator, which is linear in θ. For $\theta = 0$, it becomes $-e^{\frac{1}{\lambda}}(1-\xi) < 0$ and hence (with the minus sign at the beginning) (5) is positive. For $\theta = 1$, it becomes

$$e^{\frac{1}{\lambda}}(1 + \xi - 3\xi + 2\mu\xi - 1) = -2\xi(1-\mu) < 0$$

which is also negative. Due to linearity in θ, the (5) is always positive. Therefore ΔA is increasing in λ with a higher rate of increase for $\lambda > \widehat{\lambda}_1$ than $\widehat{\lambda} < \lambda < \widehat{\lambda}_1$.

Proof of Theorem 2 Expected cognitive cost can be rewritten as

$$\begin{aligned} C(\xi) &= C_h^\star\left(\mu^{+\hbar}\right)\mathbb{P}\{s=+, e=\hbar\} + C_h^\star\left(\mu^{+\ell}\right)\mathbb{P}\{s=+, e=\ell\} \\ &\quad + C_h^\star\left(\mu^{-\hbar}\right)\mathbb{P}\{s=-, e=\hbar\} + C_h^\star\left(\mu^{-\ell}\right)\mathbb{P}\{s=-, e=\ell\} = \mathbb{E}[\overline{H}(\mu^{se})] - \varphi(\lambda) \end{aligned} \tag{6}$$

where

$$\overline{H}(\mu) = \begin{cases} H(\mu) & \text{if } \underline{\mu} < \mu < \overline{\mu} \\ \varphi(\lambda) & \text{otherwise} \end{cases}$$

and

$$\varphi(\lambda) = \log\left(e^{1/\lambda}+1\right) - \frac{1}{\lambda}\frac{e^{1/\lambda}}{e^{1/\lambda}+1}$$

and

$$H(\mu) = -\mu \ln \mu - (1-\mu)\ln(1-\mu).$$

Note that only $\mathbb{E}[\overline{H}(\mu^{se})]$ depends on ξ. Assume now that $\xi < \underline{\xi}$ where all posteriors are in between $\underline{\mu}$ and $\overline{\mu}$. Then $\mathbb{E}[\overline{H}(\mu^{se})] = \mathbb{E}[H(\mu^{se})]$. The first order derivative of the first term in $\mathbb{E}[H(\mu^{se})]$ is

$$\begin{aligned}
&\frac{d}{d\xi}\left(H\left(\mu^{+\hbar}\right)\mathbb{P}\{s=+,e=\hbar\}\right)\\
&=\left(\frac{d}{d\xi}H\left(\mu^{+\hbar}\right)\right)\mathbb{P}\{s=+,e=\hbar\}+H\left(\mu^{+\hbar}\right)\frac{d}{d\xi}\mathbb{P}\{s=+,e=\hbar\}\\
&=\ln\left(\frac{1-\mu^{+\hbar}}{\mu^{+\hbar}}\right)\left(\frac{d}{d\xi}\mu^{+\hbar}\right)\mathbb{P}\{s=+,e=\hbar\}+H\left(\mu^{+\hbar}\right)\frac{d}{d\xi}\mathbb{P}\{s=+,e=\hbar\}\\
&=\ln\left(\frac{1-\mu^{+\hbar}}{\mu^{+\hbar}}\right)\left(\frac{1}{2}\frac{\mu\theta(1-\mu)(1-\theta)}{\xi\mu\theta+\frac{1}{2}(1-\xi)(1-\theta)}\right)+H\left(\mu^{+\hbar}\right)\left(\mu\theta-\frac{1}{2}(1-\theta)\right).
\end{aligned}$$

The second-order derivative is

$$\begin{aligned}
&\frac{d}{d\xi}\left(\ln\left(\frac{1-\mu^{+\hbar}}{\mu^{+\hbar}}\right)\left(\frac{1}{2}\frac{\mu\theta(1-\mu)(1-\theta)}{\xi\mu\theta+\frac{1}{2}(1-\xi)(1-\theta)}\right)+H\left(\mu^{+\hbar}\right)\left(\mu\theta-\frac{1}{2}(1-\theta)\right)\right)\\
&=\left(\frac{d}{d\xi}\ln\left(\frac{1-\mu^{+\hbar}}{\mu^{+\hbar}}\right)\right)\left(\frac{1}{2}\frac{\mu\theta(1-\mu)(1-\theta)}{\xi\mu\theta+\frac{1}{2}(1-\xi)(1-\theta)}\right)\\
&+\ln\left(\frac{1-\mu^{+\hbar}}{\mu^{+\hbar}}\right)\frac{d}{d\xi}\left(\frac{1}{2}\frac{\mu\theta(1-\mu)(1-\theta)}{\xi\mu\theta+\frac{1}{2}(1-\xi)(1-\theta)}\right)+\left(\mu\theta-\frac{1}{2}(1-\theta)\right)\frac{d}{d\xi}H\left(\mu^{+\hbar}\right)\\
&=\left(-\frac{1}{(1-\mu^{+\hbar})\mu^{+\hbar}}\right)\left(\frac{1}{2}\frac{\mu\theta(1-\mu)(1-\theta)}{\left(\xi\mu\theta+\frac{1}{2}(1-\xi)(1-\theta)\right)^2}\right)\left(\frac{1}{2}\frac{\mu\theta(1-\mu)(1-\theta)}{\xi\mu\theta+\frac{1}{2}(1-\xi)(1-\theta)}\right)\\
&-\ln\left(\frac{1-\mu^{+\hbar}}{\mu^{+\hbar}}\right)\frac{1}{2}\frac{(\mu\theta(1-\mu)(1-\theta))\left(\mu\theta-\frac{1}{2}(1-\theta)\right)}{\left(\xi\mu\theta+\frac{1}{2}(1-\xi)(1-\theta)\right)^2}\\
&+\left(\mu\theta-\frac{1}{2}(1-\theta)\right)\ln\left(\frac{1-\mu^{+\hbar}}{\mu^{+\hbar}}\right)\left(\frac{1}{2}\frac{\mu\theta(1-\mu)(1-\theta)}{\left(\xi\mu\theta+\frac{1}{2}(1-\xi)(1-\theta)\right)^2}\right)\\
&=-\frac{1}{4}\frac{1}{(1-\mu^{+\hbar})\mu^{+\hbar}}\frac{(\mu\theta(1-\mu)(1-\theta))^2}{\left(\xi\mu\theta+\frac{1}{2}(1-\xi)(1-\theta)\right)^3}<0.
\end{aligned}$$

That is $H\left(\mu^{+\hbar}\right)\mathbb{P}\{s=+,e=\hbar\}$ is concave in ξ. In a similar fashion, the first-order derivative of $H\left(\mu^{+\ell}\right)\mathbb{P}\{s=+,e=\ell\}$ is

$$\begin{aligned}
&\frac{d}{d\xi}\left(H\left(\mu^{+\ell}\right)\mathbb{P}\{s=+,e=\ell\}\right)\\
&=\left(\frac{d}{d\xi}H\left(\mu^{+\ell}\right)\right)\mathbb{P}\{s=+,e=\ell\}+H\left(\mu^{+\ell}\right)\frac{d}{d\xi}\mathbb{P}\{s=+,e=\ell\}\\
&=\ln\left(\frac{1-\mu^{+\ell}}{\mu^{+\ell}}\right)\left(\frac{d}{d\xi}\mu^{+\ell}\right)\mathbb{P}\{s=+,e=\ell\}+H\left(\mu^{+\ell}\right)\frac{d}{d\xi}\mathbb{P}\{s=+,e=\ell\}\\
&=\ln\left(\frac{1-\mu^{+\ell}}{\mu^{+\ell}}\right)\left(-\frac{1}{2}\frac{\mu\theta(1-\mu)(1-\theta)}{\left((1-\xi)\mu\theta+\frac{1}{2}\xi(1-\theta)\right)}\right)+H\left(\mu^{+\ell}\right)\left(-\mu\theta+\frac{1}{2}(1-\theta)\right).
\end{aligned}$$

The second-order derivative is

$$\frac{d}{d\xi}\left(\ln\left(\frac{1-\mu^{+\ell}}{\mu^{+\ell}}\right)\left(-\frac{1}{2}\frac{\mu\theta(1-\mu)(1-\theta)}{\left((1-\xi)\mu\theta+\frac{1}{2}\xi(1-\theta)\right)}\right)+H\left(\mu^{+\ell}\right)\left(-\mu\theta+\frac{1}{2}(1-\theta)\right)\right)$$

$$
\begin{aligned}
&= \left(-\frac{1}{(1-\mu^{+\ell})\mu^{+\ell}}\right)\left(\frac{d}{d\xi}\mu^{+\ell}\right)\left(-\frac{1}{2}\frac{\mu\theta\,(1-\mu)\,(1-\theta)}{\left((1-\xi)\,\mu\theta+\frac{1}{2}\xi\,(1-\theta)\right)}\right) \\
&+ \ln\left(\frac{1-\mu^{+\ell}}{\mu^{+\ell}}\right)\frac{d}{d\xi}\left(-\frac{1}{2}\frac{\mu\theta\,(1-\mu)\,(1-\theta)}{\left((1-\xi)\,\mu\theta+\frac{1}{2}\xi\,(1-\theta)\right)}\right)+\left(-\mu\theta+\frac{1}{2}\,(1-\theta)\right)\frac{d}{d\xi}H\left(\mu^{+\ell}\right) \\
&= \left(-\frac{1}{(1-\mu^{+\ell})\mu^{+\ell}}\right)\left(-\frac{1}{2}\frac{\mu\theta\,(1-\mu)\,(1-\theta)}{\left((1-\xi)\,\mu\theta+\frac{1}{2}\xi\,(1-\theta)\right)^2}\right)\left(-\frac{1}{2}\frac{\mu\theta\,(1-\mu)\,(1-\theta)}{\left((1-\xi)\,\mu\theta+\frac{1}{2}\xi\,(1-\theta)\right)}\right) \\
&+ \ln\left(\frac{1-\mu^{+\ell}}{\mu^{+\ell}}\right)\left(\frac{1}{2}\frac{\left(\mu\theta\,(1-\mu)\,(1-\theta)\right)\left(-\mu\theta+\frac{1}{2}\,(1-\theta)\right)}{\left((1-\xi)\,\mu\theta+\frac{1}{2}\xi\,(1-\theta)\right)^2}\right) \\
&+ \left(-\mu\theta+\frac{1}{2}\,(1-\theta)\right)\ln\left(\frac{1-\mu^{+\ell}}{\mu^{+\ell}}\right)\left(-\frac{1}{2}\frac{\mu\theta\,(1-\mu)\,(1-\theta)}{\left((1-\xi)\,\mu\theta+\frac{1}{2}\xi\,(1-\theta)\right)^2}\right) \\
&= -\frac{1}{4}\frac{1}{(1-\mu^{+\ell})\mu^{+\ell}}\frac{\left(\mu\theta\,(1-\mu)\,(1-\theta)\right)^2}{\left((1-\xi)\,\mu\theta+\frac{1}{2}\xi\,(1-\theta)\right)^3}<0.
\end{aligned}
$$

Thus $H\left(\mu^{+\ell}\right)\mathbb{P}\{s=+,e=\ell\}$ is also concave in ξ. Similarly, the first order derivative of $H\left(\mu^{-\hbar}\right)\mathbb{P}\{s=-,e=\hbar\}$ is

$$
\begin{aligned}
&\frac{d}{d\xi}\left(H\left(\mu^{-\hbar}\right)\mathbb{P}\{s=-,e=\hbar\}\right) \\
&= \left(\frac{d}{d\xi}H\left(\mu^{-\hbar}\right)\right)\mathbb{P}\{s=-,e=\hbar\}+H\left(\mu^{-\hbar}\right)\frac{d}{d\xi}\mathbb{P}\{s=-,e=\hbar\} \\
&= \ln\left(\frac{1-\mu^{-\hbar}}{\mu^{-\hbar}}\right)\left(\frac{d}{d\xi}\mu^{-\hbar}\right)\mathbb{P}\{s=-,e=\hbar\}+H\left(\mu^{-\hbar}\right)\frac{d}{d\xi}\mathbb{P}\{s=-,e=\hbar\} \\
&= \ln\left(\frac{1-\mu^{-\hbar}}{\mu^{-\hbar}}\right)\left(-\frac{1}{2}\frac{\mu\theta\,(1-\mu)\,(1-\theta)}{\left(\xi\,(1-\mu)\,\theta+\frac{1}{2}\,(1-\xi)\,(1-\theta)\right)}\right)+H\left(\mu^{-\hbar}\right)\left((1-\mu)\,\theta-\frac{1}{2}\,(1-\theta)\right).
\end{aligned}
$$

The second-order derivative is

$$
\begin{aligned}
&\frac{d}{d\xi}\left(\ln\left(\frac{1-\mu^{-\hbar}}{\mu^{-\hbar}}\right)\left(-\frac{1}{2}\frac{\mu\theta\,(1-\mu)\,(1-\theta)}{\left(\xi\,(1-\mu)\,\theta+\frac{1}{2}\,(1-\xi)\,(1-\theta)\right)}\right)+H\left(\mu^{-\hbar}\right)\left((1-\mu)\,\theta-\frac{1}{2}\,(1-\theta)\right)\right) \\
&= \left(\frac{d}{d\xi}\ln\frac{1-\mu^{-\hbar}}{\mu^{-\hbar}}\right)\left(-\frac{1}{2}\frac{\mu\theta\,(1-\mu)\,(1-\theta)}{\left(\xi\,(1-\mu)\,\theta+\frac{1}{2}\,(1-\xi)\,(1-\theta)\right)}\right) \\
&+ \left(\ln\frac{1-\mu^{-\hbar}}{\mu^{-\hbar}}\right)\frac{d}{d\xi}\left(-\frac{1}{2}\frac{\mu\theta\,(1-\mu)\,(1-\theta)}{\left(\xi\,(1-\mu)\,\theta+\frac{1}{2}\,(1-\xi)\,(1-\theta)\right)}\right)+\left((1-\mu)\,\theta-\frac{1}{2}\,(1-\theta)\right)\frac{d}{d\xi}H\left(\mu^{-\hbar}\right) \\
&= \left(-\frac{1}{(1-\mu^{-\hbar})\mu^{-\hbar}}\right)\left(-\frac{1}{2}\frac{\mu\theta\,(1-\mu)\,(1-\theta)}{\left(\xi\,(1-\mu)\,\theta+\frac{1}{2}\,(1-\xi)\,(1-\theta)\right)^2}\right)\left(-\frac{1}{2}\frac{\mu\theta\,(1-\mu)\,(1-\theta)}{\left(\xi\,(1-\mu)\,\theta+\frac{1}{2}\,(1-\xi)\,(1-\theta)\right)}\right) \\
&+ \left(\ln\frac{1-\mu^{-\hbar}}{\mu^{-\hbar}}\right)\left(\frac{1}{2}\frac{\left(\mu\theta\,(1-\mu)\,(1-\theta)\right)\left((1-\mu)\,\theta-\frac{1}{2}\,(1-\theta)\right)}{\left(\xi\,(1-\mu)\,\theta+\frac{1}{2}\,(1-\xi)\,(1-\theta)\right)^2}\right) \\
&+ \left((1-\mu)\,\theta-\frac{1}{2}\,(1-\theta)\right)\ln\left(\frac{1-\mu^{-\hbar}}{\mu^{-\hbar}}\right)\left(-\frac{1}{2}\frac{\mu\theta\,(1-\mu)\,(1-\theta)}{\left(\xi\,(1-\mu)\,\theta+\frac{1}{2}\,(1-\xi)\,(1-\theta)\right)^2}\right) \\
&= -\frac{1}{4}\frac{1}{(1-\mu^{-\hbar})\mu^{-\hbar}}\frac{\left(\mu\theta\,(1-\mu)\,(1-\theta)\right)^2}{\left(\xi\,(1-\mu)\,\theta+\frac{1}{2}\,(1-\xi)\,(1-\theta)\right)^3}<0.
\end{aligned}
$$

Thus $H\left(\mu^{-\hbar}\right)\mathbb{P}\{s=-,e=\hbar\}$ is also concave. Finally, the first order derivative of $H\left(\mu^{-\ell}\right)\mathbb{P}\{s=-,e=\ell\}$ is

$$
\frac{d}{d\xi}\left(H\left(\mu^{-\ell}\right)\mathbb{P}\{s=-,e=\ell\}\right)
$$

$$
\begin{aligned}
&= \left(\frac{d}{d\xi} H\left(\mu^{-\ell}\right)\right) \mathbb{P}\left\{s=-, e=\ell\right\} + H\left(\mu^{-\ell}\right) \frac{d}{d\xi} \mathbb{P}\left\{s=-, e=\ell\right\} \\
&= \ln\left(\frac{1-\mu^{-\ell}}{\mu^{-\ell}}\right)\left(\frac{d}{d\xi}\mu^{-\ell}\right) \mathbb{P}\left\{s=-, e=\ell\right\} + H\left(\mu^{-\ell}\right) \frac{d}{d\xi} \mathbb{P}\left\{s=-, e=\ell\right\} \\
&= \ln\left(\frac{1-\mu^{-\ell}}{\mu^{-\ell}}\right) \frac{1}{2} \frac{\mu\theta(1-\mu)(1-\theta)}{\left((1-\xi)(1-\mu)\theta + \frac{1}{2}\xi(1-\theta)\right)} + H\left(\mu^{-\ell}\right)\left(-(1-\mu)\theta + \frac{1}{2}(1-\theta)\right).
\end{aligned}
$$

The second-order derivative is

$$
\begin{aligned}
&\frac{d}{d\xi}\left(\ln\left(\frac{1-\mu^{-\ell}}{\mu^{-\ell}}\right) \frac{1}{2} \frac{\mu\theta(1-\mu)(1-\theta)}{\left((1-\xi)(1-\mu)\theta + \frac{1}{2}\xi(1-\theta)\right)} + H\left(\mu^{-\ell}\right)\left(-(1-\mu)\theta + \frac{1}{2}(1-\theta)\right)\right) \\
&= \left(\frac{d}{d\xi} \ln\frac{1-\mu^{-\ell}}{\mu^{-\ell}}\right) \frac{1}{2} \frac{\mu\theta(1-\mu)(1-\theta)}{\left((1-\xi)(1-\mu)\theta + \frac{1}{2}\xi(1-\theta)\right)} \\
&+ \ln\frac{1-\mu^{-\ell}}{\mu^{-\ell}} \frac{d}{d\xi}\left(\frac{1}{2} \frac{\mu\theta(1-\mu)(1-\theta)}{\left((1-\xi)(1-\mu)\theta + \frac{1}{2}\xi(1-\theta)\right)}\right) + \left(-(1-\mu)\theta + \frac{1}{2}(1-\theta)\right) \frac{d}{d\xi} H\left(\mu^{-\ell}\right) \\
&= \left(-\frac{1}{(1-\mu^{-\ell})\mu^{-\ell}}\right)\left(\frac{1}{2} \frac{\mu\theta(1-\mu)(1-\theta)}{\left((1-\xi)(1-\mu)\theta + \frac{1}{2}\xi(1-\theta)\right)^2}\right) \frac{1}{2} \frac{\mu\theta(1-\mu)(1-\theta)}{\left((1-\xi)(1-\mu)\theta + \frac{1}{2}\xi(1-\theta)\right)} \\
&- \ln\frac{1-\mu^{-\ell}}{\mu^{-\ell}}\left(\frac{1}{2}\left(\mu\theta(1-\mu)(1-\theta)\right) \frac{\left(-(1-\mu)\theta + \frac{1}{2}(1-\theta)\right)}{\left((1-\xi)(1-\mu)\theta + \frac{1}{2}\xi(1-\theta)\right)^2}\right) \\
&+ \left(-(1-\mu)\theta + \frac{1}{2}(1-\theta)\right) \ln\left(\frac{1-\mu^{-\ell}}{\mu^{-\ell}}\right)\left(\frac{1}{2} \frac{\mu\theta(1-\mu)(1-\theta)}{\left((1-\xi)(1-\mu)\theta + \frac{1}{2}\xi(1-\theta)\right)^2}\right) \\
&= -\frac{1}{4} \frac{1}{(1-\mu^{-\ell})\mu^{-\ell}} \frac{\left(\mu\theta(1-\mu)(1-\theta)\right)^2}{\left((1-\xi)(1-\mu)\theta + \frac{1}{2}\xi(1-\theta)\right)^3} < 0
\end{aligned}
$$

which means that $H\left(\mu^{-\ell}\right)\mathbb{P}\left\{s=-, e=\ell\right\}$ is also concave. This shows that all terms in $\mathbb{E}_{se}[H(\mu^{se})]$ are concave.

Assume that $\mu > 0.5$. For $\xi \in [\underline{\xi}, \overline{\xi}]$ $\mu^{+\ell}$ is outside of the search interval whose contributing term will be $\overline{H}\left(\mu^{+\hbar}\right) = \varphi(\lambda) P\left\{s=+, e=\hbar\right\}$ which is linear in ξ. Similarly, for $\xi > \underline{\xi}$, $\mu^{-\hbar}$ is also outside of the search interval whose contributing term $\overline{H}\left(\mu^{-\hbar}\right) = \varphi(\lambda) P\left\{s=-, e=\hbar\right\}$ is linear in ξ. This means that in both of these cases, $C(\xi)$ is also strictly concave in ξ.

Assuming that all posteriors are within the search interval, the first-order derivatives of the four terms in the expected cognitive effort function are

$$
\begin{aligned}
\frac{d}{d\xi}\left(H\left(\mu^{+\hbar}\right)\mathbb{P}\left\{s=+, e=\hbar\right\}\right) &= \ln\left(\frac{1-\mu^{+\hbar}}{\mu^{+\hbar}}\right) \frac{1}{2} \frac{\mu\theta(1-\mu)(1-\theta)}{\xi\mu\theta + \frac{1}{2}(1-\xi)(1-\theta)} + H\left(\mu^{+\hbar}\right)\left(\mu\theta - \frac{1}{2}(1-\theta)\right) \\
\frac{d}{d\xi}\left(H\left(\mu^{+\ell}\right)\mathbb{P}\left\{s=+, e=\ell\right\}\right) &= -\ln\left(\frac{1-\mu^{+\ell}}{\mu^{+\ell}}\right) \frac{1}{2} \frac{\mu\theta(1-\mu)(1-\theta)}{\left((1-\xi)\mu\theta + \frac{1}{2}\xi(1-\theta)\right)} \\
&\quad - H\left(\mu^{+\ell}\right)\left(\mu\theta - \frac{1}{2}(1-\theta)\right) \\
\frac{d}{d\xi}\left(H\left(\mu^{-\hbar}\right)\mathbb{P}\left\{s=-, e=\hbar\right\}\right) &= -\ln\left(\frac{1-\mu^{-\hbar}}{\mu^{-\hbar}}\right) \frac{1}{2} \frac{\mu\theta(1-\mu)(1-\theta)}{\left(\xi(1-\mu)\theta + \frac{1}{2}(1-\xi)(1-\theta)\right)} \\
&\quad - H\left(\mu^{-\hbar}\right)\left(-(1-\mu)\theta + \frac{1}{2}(1-\theta)\right) \\
\frac{d}{d\xi}\left(H\left(\mu^{-\ell}\right)\mathbb{P}\left\{s=-, e=\ell\right\}\right) &= \ln\left(\frac{1-\mu^{-\ell}}{\mu^{-\ell}}\right) \frac{1}{2} \frac{\mu\theta(1-\mu)(1-\theta)}{\left((1-\xi)(1-\mu)\theta + \frac{1}{2}\xi(1-\theta)\right)} \\
&\quad + H\left(\mu^{-\ell}\right)\left(-(1-\mu)\theta + \frac{1}{2}(1-\theta)\right).
\end{aligned}
$$

Note that at $\xi = 1/2$, we have

$$\mu^{+\hbar} = \mu^{+\ell} = \frac{\frac{1}{2}\mu(1-\theta)+\mu\theta}{\mu\theta+\frac{1}{2}(1-\theta)}$$
$$\mu^{-\hbar} = \mu^{-\ell} = \frac{\frac{1}{2}\mu(1-\theta)}{(1-\mu)\theta+\frac{1}{2}(1-\theta)}$$

and

$$\mathbb{P}\{s=+,e=\hbar\} = \mathbb{P}\{s=+,e=\ell\} = \mu\theta+\frac{1}{2}(1-\theta)$$
$$\mathbb{P}\{s=-,e=\hbar\} = \mathbb{P}\{s=-,e=\ell\} = (1-\mu)\theta+\frac{1}{2}(1-\theta)$$

Then, we also have

$$\frac{d}{d\xi}\left(H\left(\mu^{+\hbar}\right)\mathbb{P}\{s=+,e=\hbar\}\right) = -\frac{d}{d\xi}\left(H\left(\mu^{+\ell}\right)\mathbb{P}\{s=+,e=\ell\}\right)$$

and

$$\frac{d}{d\xi}\left(H\left(\mu^{-\hbar}\right)\mathbb{P}\{s=-,e=\hbar\}\right) = -\frac{d}{d\xi}\left(H\left(\mu^{-\ell}\right)\mathbb{P}\{s=-,e=\ell\}\right).$$

Then this means that at $\xi = 1/2$, we have

$$\frac{d}{d\xi}\mathbb{E}[\overline{H}(\mu^{se})] = 0.$$

Since $\mathbb{E}[\overline{H}(\mu^{se})]$ is strictly concave, $\xi = 1/2$ is the maximizer of $\mathbb{E}_{se}[\overline{H}(\mu^{se})]$ and it is decreasing for $\xi \in [1/2, \underline{\xi})$, i.e., when all posteriors are in the search interval.

Assume now that $\xi > \overline{\xi}$. In this case, we only have $\mu^{+\ell}$ and $\mu^{-\ell}$ within the search range since both $\mu^{+\ell}$ and $\mu^{-\ell}$ approach to μ as $\xi \uparrow 1$. Furthermore, at $\xi = 1$, we have

$$\mathbb{P}\{s=+,e=\ell\} = \mathbb{P}\{s=-,e=\ell\} = \frac{1}{2}(1-\theta).$$

Then, the first-order derivative of $\mathbb{E}[\overline{H}(\mu^{se})]$ at $\xi = 1$ is

$$\begin{aligned}
&\frac{d}{d\xi}\left(H\left(\mu^{+\ell}\right)\mathbb{P}\{s=+,e=\ell\}\right)_{|\xi=1} + \frac{d}{d\xi}\left(H\left(\mu^{-\ell}\right)\mathbb{P}\{s=-,e=\ell\}\right)_{|\xi=1} \\
&+\varphi(\lambda)\frac{d}{d\xi}\mathbb{P}\{s=+,e=\hbar\} + \varphi(\lambda)\frac{d}{d\xi}\mathbb{P}\{s=-,e=\hbar\} \\
&= -\ln\left(\frac{1-\mu}{\mu}\right)\mu\theta(1-\mu) - H(\mu)\left(\mu\theta - \frac{1}{2}(1-\theta)\right) + \ln\left(\frac{1-\mu}{\mu}\right)\mu\theta(1-\mu) \\
&+ H(\mu)\left(-(1-\mu)\theta + \frac{1}{2}(1-\theta)\right) \\
&+\varphi(\lambda)\left(\mu\theta - \frac{1}{2}(1-\theta) + (1-\mu)\theta - \frac{1}{2}(1-\theta)\right) \\
&= (H(\mu) - \varphi(\lambda))(1-2\theta).
\end{aligned}$$

Since $H(\mu) > \varphi(\lambda)$ (as μ is in the search interval), the first-order derivative of $\mathbb{E}[\overline{H}(\mu^{se})]$ at $\xi = 1$ is positive (resp. negative) when $\theta < 0.5$ (resp. $\theta > 0.5$). This means that in the last region, i.e., $\xi \in [\overline{\xi}, 1]$, the expected cognitive effort is increasing in ξ if $\theta \leq 0.5$, and either decreasing or unimodal in ξ when $\theta > 0.5$ due to strict concavity.

Proof of Proposition 4 Note that at $\lambda = 0$, both belief thresholds are at the extremes, and the DM always processes information as it is free. As λ increases, $\overline{\mu}$ and $\underline{\mu}$ start to move towards 0.5. This means that there exists a threshold λ_1 such that for all $\lambda < \lambda_1$, all posterior beliefs are in the search region $\left[\underline{\mu}, \overline{\mu}\right]$. Assume that $\lambda < \lambda_1$. Then,

$$\Delta C = C(\xi) - C_{ne} = \lambda \left(\sum_{(s,e)\in\{+,-\}\times\{\hbar,\ell\}} \mathbb{P}\{s,e\}(H(\mu^{se}) - \varphi(\lambda)) - \sum_{s\in\{+,-\}} \mathbb{P}\{s\}(H(\mu^{s}) - \varphi(\lambda)) \right)$$
$$= \lambda \left(\sum_{(s,e)\in\{+,-\}\times\{\hbar,\ell\}} \mathbb{P}\{s,e\} H(\mu^{se}) - \sum_{s\in\{+,-\}} \mathbb{P}\{s\} H(\mu^{s}) \right). \tag{7}$$

Note that

$$\mathbb{P}\{-,\hbar\} H\left(\mu^{-\hbar}\right) + \mathbb{P}\{-,\ell\} H\left(\mu^{-\ell}\right) \le \mathbb{P}\{s=-\} H(\mu^{-})$$

due to Jensen's inequality since

$$\mathbb{P}\{e=\hbar|s=-\} H\left(\mu^{-\hbar}\right) + \mathbb{P}\{e=\ell|s=-\} H\left(\mu^{-\ell}\right) \le H(\mu^{-})$$

and H is concave. Similarly,

$$\mathbb{P}\{+,\hbar\} H\left(\mu^{+\hbar}\right) + \mathbb{P}\{+,\ell\} H\left(\mu^{+\ell}\right) \le \mathbb{P}\{s=+\} H\left(\mu^{+}\right).$$

This means that (7), i.e., ΔC, is negative and linearly decreasing. Assume $\mu > 0.5$. Then λ_1 precisely corresponds to the information cost level where $\mu^{+\hbar} = \overline{\mu}$. Furthermore, there exists another threshold $\lambda_2 > \lambda_1$ such that $\mu^{-\hbar} = \underline{\mu}$. (Due to the symmetry of cognitive effort with respect to μ, for $\mu < 0.5$, only the order changes, and the rest of the analysis stays same). Assume that $\lambda_1 < \lambda < \lambda_2$. Then

$$\Delta C = \lambda \left(\sum_{(s,e)\in\{(+,\ell),(-,\hbar),(-,\ell)\}} \mathbb{P}\{s,e\}(H(\mu^{se}) - \varphi(\lambda)) - \sum_{s\in\{+,-\}} \mathbb{P}\{s\}(H(\mu^{s}) - \varphi(\lambda)) \right)$$
$$= \lambda \left(\sum_{(s,e)\in\{(+,\ell),(-,\hbar),(-,\ell)\}} \mathbb{P}\{s,e\} H(\mu^{se}) - \sum_{s\in\{+,-\}} \mathbb{P}\{s\} H(\mu^{s}) \right) + \mathbb{P}\{+,\hbar\}\lambda\varphi(\lambda).$$

The first-order derivative is

$$\frac{\partial}{\partial\lambda}\Delta C = \sum_{(s,e)\in\{(+,\ell),(-,\hbar),(-,\ell)\}} \mathbb{P}\{s,e\} H(\mu^{se}) - \sum_{s\in\{+,-\}} \mathbb{P}\{s\} H(\mu^{s}) + \mathbb{P}\{+,\hbar\} \frac{\partial}{\partial\lambda}\lambda\varphi(\lambda).$$

$\lambda\varphi(\lambda)$ is increasing in λ and $\frac{\partial}{\partial\lambda}\lambda\varphi(\lambda) > 0$. In addition,

$$\sum_{(s,e)\in\{(+,\ell),(-,\hbar),(-,\ell)\}} \mathbb{P}\{s,e\} H(\mu^{se}) - \sum_{s\in\{+,-\}} \mathbb{P}\{s\} H(\mu^{s}) < 0.$$

For $\lambda > \lambda_2$, we have

$$\frac{\partial}{\partial\lambda}\Delta C = \sum_{(s,e)\in\{(+,\ell),(-,\ell)\}} \mathbb{P}\{s,e\} H(\mu^{se}) - \sum_{s\in\{+,-\}} \mathbb{P}\{s\} H(\mu^{s}) + (\mathbb{P}\{+,\hbar\} + \mathbb{P}\{-,\hbar\}) \frac{\partial}{\partial\lambda}\lambda\varphi(\lambda).$$

where

$$\sum_{(s,e)\in\{(+,\ell),(-,\ell)\}} \mathbb{P}\{s,e\} H(\mu^{se}) - \sum_{s\in\{+,-\}} \mathbb{P}\{s\} H(\mu^{s}) < 0.$$

This means that there exists a threshold $\widehat{\lambda}_\xi$ such that for $\lambda > \widehat{\lambda}_\xi$, $\Delta C > 0$ and is increasing in λ. Note that it is still possible that $\widehat{\lambda}_\xi$ can be higher than the natural bound $\overline{\lambda}$ imposed by complementarity per Proposition (2).

Proof of Theorem 3 Assume $\mu > 1/2$. We first investigate the extreme cognitive load scenario. Assume that $\lambda \uparrow \overline{\lambda}$. Then $\underline{\xi} \downarrow 1/2$ (i.e., the first interval in Theorem 1 vanishes) and $\mu^{+\hbar} = \overline{\mu}$ which gives

$$\mu^{+\hbar} = \overline{\mu} \Leftrightarrow \frac{\frac{1}{2}\mu(1-\theta)+\mu\theta}{\mu\theta+\frac{1}{2}(1-\theta)} = \frac{e^{1/\lambda}}{e^{1/\lambda}+1} \Leftrightarrow \lambda^* = \frac{1}{\ln\left(\frac{\mu(1+\theta)}{(1-\theta)(1-\mu)}\right)}.$$

To analyze the behavior of the second interval with respect to ξ, we look at the first-order derivative of the cognitive effort function. When $\lambda \uparrow \overline{\lambda}$, the cognitive effort implied by the posterior $\mu^{+\hbar}$ also vanishes. Let us look at its derivative at $\xi = 1/2$. We have

$$\begin{aligned}
&\frac{\partial}{\partial \xi}\left(H\left(\mu^{+\hbar}\right)-\varphi(\lambda)\right)\mathbb{P}\{s=+,e=\hbar\} \\
&= \ln\left(\frac{1-\mu^{+\hbar}}{\mu^{+\hbar}}\right)\frac{1}{2}\frac{\mu\theta(1-\mu)(1-\theta)}{\xi\mu\theta+\frac{1}{2}(1-\xi)(1-\theta)} + \left(H\left(\mu^{+\hbar}\right)-\varphi(\lambda)\right)\left(\mu\theta-\frac{1}{2}(1-\theta)\right) \\
&= \ln\left(\frac{1-\mu^{+\hbar}}{\mu^{+\hbar}}\right)\frac{\mu\theta(1-\mu)(1-\theta)}{\mu\theta+\frac{1}{2}(1-\theta)} + \left(H\left(\mu^{+\hbar}\right)-\varphi(\lambda)\right)\left(\mu\theta-\frac{1}{2}(1-\theta)\right).
\end{aligned} \tag{8}$$

Note that since $\mu > 0.5$ by assumption, we have $\mu^{+\hbar} > 0.5$ and hence the term inside the log term is less than 1. Therefore, the first term in (8) is negative. The second term equals to zero as $H\left(\mu^{+\hbar}\right) - \varphi(\lambda) = 0$ since $\mu^{+\hbar} = \overline{\mu}$ by assumption. Then (8) is negative for $\lambda = \overline{\lambda}$. Note that at $\xi = 1/2$, the first order derivative of the cognitive effort function is 0 and since it is concave, $\xi = 1/2$ is the maximizer in the first region (see the Proof of Theorem 2). Therefore, it must be that

$$\frac{\partial}{\partial \xi}\sum_{\{s,e\}\in(\{+,\ell\},\{-,\hbar\},\{-,\ell\})}\left(\left(H\left(\mu^{se}\right)-\varphi(\lambda)\right)\mathbb{P}\{s,e\}\right)_{|\xi=1/2} > 0.$$

Hence, the expected cognitive effort is increasing when $\xi \downarrow 1/2$ and $\lambda \uparrow \overline{\lambda}$.

Next, we show that $C(\xi)$ is continuous in λ. Note that μ^{se}, $H(\mu^{se})$ and $\mathbb{P}\{s,e\}$ are all constant in λ. The only term in $C(\xi)$ that depends on λ is the residual uncertainty function $\varphi(\lambda)$ and the endpoints $\underline{\mu}$ and $\overline{\mu}$. It is easy to show that $\varphi(\lambda)$ is continuous as $\ln\left(e^{1/\lambda}+1\right)$, $\frac{1}{\lambda}$ and $\frac{e^{1/\lambda}}{e^{1/\lambda}+1}$ are all continuous for $\lambda > 0$. Hence $\varphi(\lambda)$ is also continuous. We need to check that $C(\xi)$ is continuous at the endpoints. Note that

$$\mu = \underline{\mu} \Leftrightarrow \lambda = \ln\left(\frac{1-\mu}{\mu}\right)^{-1}.$$

At $\lambda = \ln\left(\frac{1-\mu}{\mu}\right)^{-1}$, we have

$$\begin{aligned}
\varphi(\lambda) &= \ln\left(e^{\ln\left(\frac{1-\mu}{\mu}\right)}+1\right) - \frac{1}{\ln\left(\frac{1-\mu}{\mu}\right)^{-1}}\frac{e^{\ln\left(\frac{1-\mu}{\mu}\right)}}{e^{\ln\left(\frac{1-\mu}{\mu}\right)}+1} = \ln\frac{1}{\mu} - \ln\left(\frac{1-\mu}{\mu}\right)(1-\mu) \\
&= -\ln\mu - \ln(1-\mu)(1-\mu) + \ln(\mu)(1-\mu) = -\mu\ln\mu - \ln(1-\mu)(1-\mu) = H(\mu).
\end{aligned}$$

Hence as $\lambda \to \ln\left(\frac{1-\mu}{\mu}\right)^{-1}$, $\varphi(\lambda) \to H(\mu)$ and $C_h^\star(\xi) \to 0$. Similarly, one can easily show this for the other endpoint $\overline{\mu}$. This shows that $C_h^\star(\mu^{se})\mathbb{P}\{s,e\}$ and in turn $C(\xi)$ is continuous in λ. Note that since $C(\xi)$ is continuous in λ, and is increasing in ξ as $\xi \downarrow 1/2$, there must be a threshold $\overline{\lambda}_\mu$ such that the cognitive effort is higher compared to the benchmark $C_{ne} = C(\xi = 1/2)$ for $\lambda > \overline{\lambda}_\mu$.

Additionally, let $I_\theta = [\widehat{\mu}_t, 1-\widehat{\mu}_t]$ in Proposition 2 for $\widehat{\mu}_t \in (\overline{\mu}, 0.5)$. Then note that $\lambda \uparrow \overline{\lambda} \Leftrightarrow \mu \downarrow \widehat{\mu}_t$ or $\lambda \uparrow \overline{\lambda} \Leftrightarrow \mu \uparrow 1-\widehat{\mu}_t$ for $\mu > 0.5$ and $\mu < 0.5$ respectively. We can similarly show that $C(\xi)$ is continuous in

μ. First, the posterior functions μ^{se} are all continuous in μ for $\mu \in (0,1)$ and $\theta \in (0,1)$ as they are divisions of two linear functions and their denominators are nonzero. Furthermore, the entropy function $H(\mu)$ is continuous in $\mu \in (0,1)$ since it is the sum of two continuous functions. More specifically, both $-\mu \log \mu$ and $-(1-\mu)\log(1-\mu)$ are products of a linear function and the log function which is also continuous. It is known that a function of a continuous function is also continuous, hence $H(\mu^{se})$ is continuous for $(s,e) \in \{(+,-)\times(\hbar,\ell)\}$. Since $\varphi(\lambda)$ is linear in μ, $\lambda(H(\mu^{se}) - \varphi(\lambda))$ is continuous in the search interval, i.e., when $\underline{\mu} < \mu < \overline{\mu}$. Outside of the search interval, it is zero. Therefore, as before, we just need to check the endpoints. As $\mu \to \underline{\mu}$, we have

$$
\begin{aligned}
H(\underline{\mu}) &= -\frac{1}{e^{1/\lambda}+1}\ln\frac{1}{e^{1/\lambda}+1} - \left(1-\frac{1}{e^{1/\lambda}+1}\right)\ln\left(1-\frac{1}{e^{1/\lambda}+1}\right) \\
&= \ln\left(e^{1/\lambda}+1\right) - \frac{1}{\lambda}\frac{e^{1/\lambda}}{e^{1/\lambda}+1} = \varphi(\lambda)
\end{aligned}
$$

and hence $H(\underline{\mu}) - \varphi(\lambda) \to 0$. Similarly, it is easy to show that $\lambda(H(\overline{\mu}) - \varphi(\lambda)) \to 0$ since H is symmetric around 0.5 and $\overline{\mu} = 1-\underline{\mu}$. This shows that $\max\{\lambda(H(\mu^{se}) - \varphi(\lambda)), 0\}$ is continuous since μ^{se} is continuous. As $\mathbb{P}\{s,e\}$ is linear in μ for $(s,e) \in \{(+,-)\times(\hbar,\ell)\}$, $\sum_{s,e}\max\{\lambda(H(\mu^{se}) - \varphi(\lambda)), 0\}$ is also continuous which shows that $C(\xi)$ is continuous in μ. This means that there must exists a threshold μ_λ such that for $\mu < \widetilde{\mu}_\lambda$ and $\mu > 1-\widetilde{\mu}_\lambda$, expected cognitive effort $C(\xi)$ is higher than the benchmark $C_{ne} = C(\xi = 1/2)$ since C is continuous in μ and increasing in ξ as $\xi \downarrow 1/2$.

For $\mu = 0.5$, $\underline{\xi} = \overline{\xi}$ and $C(\xi)$ is decreasing in ξ for $\xi < \underline{\xi}$. Furthermore, it is increasing in ξ for $\xi > \underline{\xi}$ when $\theta \le 1/2$ per Theorem 2. This means that for $\lambda > \widetilde{\lambda}_\mu$, there exists a threshold for explainability level $\tilde{\xi} > \underline{\xi} = \overline{\xi}$ such that $\Delta C > 0$ for $\xi \in \left[\tilde{\xi}, 1\right]$.

Lemma A.1. *For $0 \le \gamma^\ell < \gamma^\hbar < 0.5$, there exists a unique threshold $\xi_d \in (0,1)$ such that $\mu^{+\hbar} = \mu^{+\ell}$ and $\mu^{-\ell} = \mu^{-\hbar}$ when $\xi = \xi_d$, $\mu^{+\hbar} < \mu^{+\ell}$ and $\mu^{-\ell} < \mu^{-\hbar}$ when $\xi < \xi_d$ and $\mu^{+\hbar} > \mu^{+\ell}$ and $\mu^{-\ell} > \mu^{-\hbar}$ otherwise. Further, $\mu^{+e} > \mu^{-e}$ for $e \in \{\hbar, \ell\}$ and $\mu^{+\hbar}$ and $\mu^{-\ell}$ are increasing in ξ while $\mu^{-\hbar}$ and $\mu^{+\ell}$ are decreasing in ξ. For $\xi = 1$, $\mu^{+\ell} = \mu^{-\ell} = \mu$.*

Proof of Lemma A.1 To find θ^{se}, we first need to compute $\mathbb{P}\{s,e|x\}$. Given $X = \mathfrak{h}$, we have

$$
\begin{aligned}
\mathbb{P}\{s=+, e=\hbar | X=\mathfrak{h}\} &= \mathbb{P}\{s=+|e=\hbar, X=\mathfrak{h}, \omega=1\}\,\mathbb{P}\{e=\hbar|X=\mathfrak{h}, \omega=1\}\,\mathbb{P}\{\omega=1|X=\mathfrak{h}\} \\
&\quad + \mathbb{P}\{s=+|e=\hbar, X=\mathfrak{h}, \omega=0\}\,\mathbb{P}\{e=\hbar|X=\mathfrak{h}, \omega=0\}\,\mathbb{P}\{\omega=0|X=\mathfrak{h}\} \\
&= \left(1-\gamma^\hbar\right)\xi\mu + \gamma^\hbar\xi(1-\mu) \\
&= \xi\left[\left(1-\gamma^\hbar\right)\mu + \gamma^\hbar(1-\mu)\right].
\end{aligned}
$$

In a similar manner, we have

$$
\begin{aligned}
\mathbb{P}\{s=+, e=\ell | x=1\} &= (1-\xi)\left[\left(1-\gamma^\ell\right)\mu + \gamma^\ell(1-\mu)\right] \\
\mathbb{P}\{s=-, e=\hbar | x=1\} &= \xi\left[\gamma^\hbar\mu + \left(1-\gamma^\hbar\right)(1-\mu)\right] \\
\mathbb{P}\{s=-, e=\ell | x=1\} &= (1-\xi)\left[\gamma^\ell\mu + \left(1-\gamma^\ell\right)(1-\mu)\right].
\end{aligned}
$$

When $X = \mathfrak{l}$, we have

$$\mathbb{P}\{s=+, e=\hbar|X=\mathfrak{l}\} = \mathbb{P}\{s=-, e=\hbar|X=\mathfrak{l}\} = \frac{1}{2}(1-\xi)$$
$$\mathbb{P}\{s=+, e=\ell|X=\mathfrak{l}\} = \mathbb{P}\{s=-, e=\ell|X=\mathfrak{l}\} = \frac{1}{2}\xi$$

since S and E are conditionally independent given $X = \mathfrak{l}$. Then we can write

$$\begin{aligned}
\theta^{+\hbar} &= \mathbb{P}\{X=\mathfrak{l}|s=+, e=\hbar\} \\
&= \frac{\mathbb{P}\{s=+, e=\hbar|X=\mathfrak{l}\}\mathbb{P}\{X=\mathfrak{l}\}}{\mathbb{P}\{s=+, e=\hbar|X=\mathfrak{h}\}\mathbb{P}\{X=\mathfrak{h}\} + \mathbb{P}\{s=+, e=\hbar|X=\mathfrak{l}\}\mathbb{P}\{X=\mathfrak{l}\}} \\
&= \frac{\theta\xi\left[\left(1-\gamma^{\hbar}\right)\mu + \gamma^{\hbar}(1-\mu)\right]}{\theta\xi\left[(1-\gamma^{\hbar})\mu + \gamma^{\hbar}(1-\mu)\right] + (1-\theta)\frac{1}{2}(1-\xi)}.
\end{aligned}$$

In a similar manner, we obtain

$$\begin{aligned}
\theta^{+\ell} &= \frac{\theta(1-\xi)\left[\left(1-\gamma^{\ell}\right)\mu + \gamma^{\ell}(1-\mu)\right]}{\theta(1-\xi)\left[(1-\gamma^{\ell})\mu + \gamma^{\ell}(1-\mu)\right] + (1-\theta)\frac{1}{2}\xi} \\
\theta^{-\hbar} &= \frac{\theta\xi\left[\gamma^{\hbar}\mu + \left(1-\gamma^{\hbar}\right)(1-\mu)\right]}{\theta\xi\left[\gamma^{\hbar}\mu + (1-\gamma^{\hbar})(1-\mu)\right] + (1-\theta)\frac{1}{2}(1-\xi)} \\
\theta^{+\ell} &= \frac{\theta(1-\xi)\left[\gamma^{\ell}\mu + \left(1-\gamma^{\ell}\right)(1-\mu)\right]}{\theta(1-\xi)\left[\gamma^{\ell}\mu + (1-\gamma^{\ell})(1-\mu)\right] + (1-\theta)\frac{1}{2}\xi}.
\end{aligned}$$

Plugging these in, we can characterize the posterior beliefs as

$$\begin{aligned}
\mu^{+\hbar} &= \frac{\theta\xi\left(1-\gamma^{\hbar}\right)\mu + (1-\theta)\frac{1}{2}(1-\xi)\mu}{\theta\xi\left[(1-\gamma^{\hbar})\mu + \gamma^{\hbar}(1-\mu)\right] + (1-\theta)\frac{1}{2}(1-\xi)} \\
\mu^{+\ell} &= \frac{\theta(1-\xi)\left(1-\gamma^{\ell}\right)\mu + (1-\theta)\frac{1}{2}\xi\mu}{\theta(1-\xi)\left[(1-\gamma^{\ell})\mu + \gamma^{\ell}(1-\mu)\right] + (1-\theta)\frac{1}{2}\xi} \\
\mu^{-\hbar} &= \frac{\theta\xi\gamma^{\hbar}\mu + (1-\theta)\frac{1}{2}(1-\xi)\mu}{\theta\xi\left[\gamma^{\hbar}\mu + (1-\gamma^{\hbar})(1-\mu)\right] + (1-\theta)\frac{1}{2}(1-\xi)} \\
\mu^{-\ell} &= \frac{\theta(1-\xi)\gamma^{\ell}\mu + (1-\theta)\frac{1}{2}\xi\mu}{\theta(1-\xi)\left[\gamma^{\ell}\mu + (1-\gamma^{\ell})(1-\mu)\right] + (1-\theta)\frac{1}{2}\xi}.
\end{aligned}$$

Taking the first order derivatives, we obtain

$$\begin{aligned}
\frac{d}{d\xi}\mu^{+\hbar} &= \frac{2\theta\mu(1-\theta)(1-\mu)\left(1-2\gamma^{\hbar}\right)}{\left(\theta\xi - \xi - \theta + 2\gamma^{\hbar}\theta\xi + 2\theta\mu\xi - 4\gamma^{\hbar}\theta\mu\xi + 1\right)^2} > 0 \\
\frac{d}{d\xi}\mu^{+\ell} &= -\frac{2\theta\mu(1-\theta)(1-\mu)\left(1-2\gamma^{\ell}\right)}{\left(\xi + 2\gamma^{\ell}\theta + 2\theta\mu - \theta\xi - 4\gamma^{\ell}\theta\mu - 2\gamma^{\ell}\theta\xi - 2\theta\mu\xi + 4\gamma^{\ell}\theta\mu\xi\right)^2} < 0 \\
\frac{d}{d\xi}\mu^{-\hbar} &= -\frac{2\theta\mu(1-\theta)(1-\mu)\left(1-2\gamma^{\hbar}\right)}{\left(\theta + \xi - 3\theta\xi + 2\gamma^{\hbar}\theta\xi + 2\theta\mu\xi - 4\gamma^{\hbar}\theta\mu\xi - 1\right)^2} < 0 \\
\frac{d}{d\xi}\mu^{-\ell} &= \frac{2\theta\mu(1-\theta)(1-\mu)(1-2B)}{\left(2\theta + \xi - 2\gamma^{\ell}\theta - 2\theta\mu - 3\theta\xi + 4\gamma^{\ell}\theta\mu + 2\gamma^{\ell}\theta\xi + 2\theta\mu\xi - 4\gamma^{\ell}\theta\mu\xi\right)^2} > 0
\end{aligned}$$

which shows the monotonicity of posteriors with respect to ξ.

Note that at $\xi = 1/2$,

$$\begin{aligned}
\mu^{+\ell} &= \frac{\theta\left(1-\gamma^{\ell}\right)\mu + (1-\theta)\frac{1}{2}\mu}{\theta\left[(1-\gamma^{\ell})\mu + \gamma^{\ell}(1-\mu)\right] + (1-\theta)\frac{1}{2}} \\
&> \frac{\theta\left(1-\gamma^{\hbar}\right)\mu + (1-\theta)\frac{1}{2}\mu}{\theta\left[(1-\gamma^{\hbar})\mu + \gamma^{\hbar}(1-\mu)\right] + (1-\theta)\frac{1}{2}} = \mu^{+\hbar}
\end{aligned}$$

since $\gamma^{\hbar} > \gamma^{\ell}$. Similarly, at $\xi = 1/2$,

$$\mu^{-\hbar} = \frac{\theta\gamma^{\hbar}\mu + (1-\theta)\,\frac{1}{2}\mu}{\theta\left[\gamma^{\hbar}\mu + (1-\gamma^{\hbar})\,(1-\mu)\right] + (1-\theta)\,\frac{1}{2}}$$
$$> \frac{\theta\gamma^{\ell}\mu + (1-\theta)\,\frac{1}{2}\mu}{\theta\left[\gamma^{\ell}\mu + (1-\gamma^{\ell})\,(1-\mu)\right] + (1-\theta)\,\frac{1}{2}} = \mu^{-\ell}.$$

In contrast, at $\xi = 1$,

$$\mu^{+\hbar} = \frac{\theta\left(1-\gamma^{\hbar}\right)\mu}{\theta\left[(1-\gamma^{\hbar})\,\mu + \gamma^{\hbar}\,(1-\mu)\right]} > \mu = \mu^{+\ell}$$

and

$$\mu^{-\hbar} = \frac{\theta\gamma^{\hbar}\mu}{\theta\left[\gamma^{\hbar}\mu + (1-\gamma^{\hbar})\,(1-\mu)\right]} < \mu = \mu^{-\ell}.$$

Since posteriors are monotone, there must be an explainability level where $\mu^{+\hbar} = \mu^{+\ell}$ and $\mu^{-\hbar} = \mu^{-\ell}$.

Proof of Proposition 5 Overreliance is

$$O(\xi) = \mathbb{P}\{s=+\}\mathbb{E}[\alpha\left(\mu^{+E}\right)|s=+] + \mathbb{P}\{s=-\}\mathbb{E}[\beta\left(\mu^{-E}\right)|s=-].$$

When $\xi < \underline{\xi}$, all posterior beliefs are in the search interval and therefore false-positive and false-negative functions α and β are linear in posterior belief. Then overreliance becomes

$$O(\xi) = \mathbb{P}\{s=+\}\alpha(\mathbb{E}[\mu^{+E}|s=+]) + \mathbb{P}\{s=-\}\beta(\mathbb{E}[\mu^{-E}|s=-])$$
$$= \mathbb{P}\{s=+\}\alpha\left(\mu^{+}\right) + \mathbb{P}\{s=-\}\beta\left(\mu^{-}\right)$$

Posterior belief μ^{+} can be computed as

$$\mu^{+} = \mathbb{P}\{\omega=1|s=+\} = \frac{\mathbb{P}\{s=+|\omega=1\}\mu}{\mathbb{P}\{s=+\}}$$
$$= \frac{\left[\xi\left(\gamma^{\ell}-\gamma^{\hbar}\right)+1-\gamma^{\ell}\right]\theta + \frac{1}{2}\,(1-\theta)}{\mathbb{P}\{s=+\}}\mu$$

where

$$\mathbb{P}\{s=+|\omega=1\} = \mathbb{P}\{s=+|\omega=1, X=\mathfrak{h}\}\mathbb{P}\{X=\mathfrak{h}|\omega=1\}$$
$$+ \mathbb{P}\{s=+|\omega=1, X=\mathfrak{l}\}\mathbb{P}\{X=\mathfrak{l}|\omega=1\}$$
$$= \left[\xi\left(\gamma^{\ell}-\gamma^{\hbar}\right)+1-\gamma^{\ell}\right]\theta + \frac{1}{2}\,(1-\theta)\,.$$

Similarly, we have

$$\mu^{-} = \mathbb{P}\{\omega=1|s=-\} = \frac{\mathbb{P}\{s=-|\omega=1\}}{\mathbb{P}\{s=-\}}\mu$$
$$= \frac{\left[\gamma^{\ell}-\xi\left(\gamma^{\ell}-\gamma^{\hbar}\right)\right]\theta + \frac{1}{2}\,(1-\theta)}{\mathbb{P}\{s=-\}}\mu.$$

Then

$$O\left(\xi\right) = \frac{\frac{\left[\xi\left(\gamma^{\ell}-\gamma^{\hbar}\right)+1-\gamma^{\ell}\right]\theta+\frac{1}{2}(1-\theta)}{\mathbb{P}\{s=+\}}\mu\left(e^{1/\lambda}+1\right)-1}{e^{2/\lambda}-1}\mathbb{P}\{s=+\}$$
$$+ \frac{e^{1/\lambda} - \frac{\left[\gamma^{\ell}-\xi\left(\gamma^{\ell}-\gamma^{\hbar}\right)\right]\theta+\frac{1}{2}(1-\theta)}{\mathbb{P}\{s=-\}}\mu\left(e^{1/\lambda}+1\right)}{e^{2/\lambda}-1}\mathbb{P}\{s=-\}$$

Taking the first order derivative with respect to ξ, we obtain

$$\begin{aligned}\frac{d}{d\xi}O(\xi) &= \frac{2\left(\gamma^{\hbar}-\gamma^{\ell}\right)\theta\mu - \frac{d\mathbb{P}\{S=-\}}{d\xi}}{1-e^{1/\lambda}}\\ &= \theta\frac{2\mu\left(\gamma^{\hbar}-\gamma^{\ell}\right)-\left(\gamma^{\ell}-\gamma^{\hbar}\right)(1-2\mu)}{1-e^{1/\lambda}}\\ &= -\theta\frac{\gamma^{\ell}-\gamma^{\hbar}}{1-e^{1/\lambda}}<0\end{aligned}$$

where

$$\begin{aligned}\frac{d\mathbb{P}\{s=-\}}{d\xi} &= \frac{d}{d\xi}\left(\left[-\xi\left(\gamma^{\ell}-\gamma^{\hbar}\right)+\gamma^{\ell}\right]\theta+\frac{1}{2}(1-\theta)\right)\mu\\ &\quad+\frac{d}{d\xi}\left(\left[\xi\left(\gamma^{\ell}-\gamma^{\hbar}\right)+1-\gamma^{\ell}\right]\theta+\frac{1}{2}(1-\theta)\right)(1-\mu)\\ &= \frac{d}{d\xi}\left(\left[-\xi\left(\gamma^{\ell}-\gamma^{\hbar}\right)+\gamma^{\ell}\right]\theta\mu+\left[\xi\left(\gamma^{\ell}-\gamma^{\hbar}\right)+1-\gamma^{\ell}\right]\theta(1-\mu)\right)\\ &= \left(\gamma^{\ell}-\gamma^{\hbar}\right)\theta(1-2\mu).\end{aligned}$$

This means that overreliance decreases in ξ when all posteriors are in the search interval and $\gamma^{\hbar}>\gamma^{\ell}$ (i.e., when there is a negative (resp. positive) relation between accuracy and explainability) since $\frac{d}{d\xi}O(\xi)<0$. Overreliance is constant when $\gamma^{\hbar}=\gamma^{\ell}$, same as in our baseline model.

Assume now that $\mu>0.5$ and $\mu^{+\hbar}\geq\overline{\mu}$. Then $\xi>\underline{\xi}$. Note that the false-positive function α is linearly increasing when $\mu^{+\hbar}<\overline{\mu}$, and linearly decreasing when $\mu^{+\hbar}\geq\overline{\mu}$. Therefore, overreliance starts to decrease in ξ with a higher rate when $\xi>\underline{\xi}$ since $\mu^{+\hbar}$ increases in ξ per Lemma A.1. In a similar manner, when $\mu^{-\hbar}\leq\underline{\mu}$, false-negative function β is linearly increasing in $\mu^{-\hbar}$. However, since $\mu^{-\hbar}$ decreases in ξ as per Lemma A.1, overreliance starts to decrease with a higher rate as ξ increases. Therefore, $O(\xi)$ is concave decreasing. The same logic applies when $\mu<0.5$, as posteriors are symmetric around 0.5, as well as for the underreliance, as functions α and β are also symmetric around 1/2.

For accuracy $O(\xi)=\mathbb{E}\left[A\left(\mu^{SE}\right)\right]$, we use a similar approach. Note that when all posteriors are within the search region, the accuracy function is linear and constant, and hence its expectation, such that

$$\mathbb{E}\left[A\left(\mu^{SE}\right)\right]=A\left(\mathbb{E}\left[\mu^{SE}\right]\right)=A(\mu)=\frac{e^{1/\lambda}}{e^{1/\lambda}+1}.$$

Assuming again $\mu>1/2$, when $\mu^{+\hbar}\geq\overline{\mu}$, accuracy function is increasing for $\mu^{+\hbar}$ which is also increasing in ξ. Therefore, accuracy increases. Similarly, when $\mu^{-\hbar}\leq\underline{\mu}$, accuracy function is decreasing in $\mu^{-\hbar}$ which is also decreasing in ξ. Therefore, accuracy starts to increase at a higher rate.

Proof of Proposition 6 Assume that $\xi<\underline{\xi}$ and all posteriors are within the search interval. We take the second-order derivative of each term of $C(\xi)=\mathbb{E}\left[C^{\star}_{\hbar}\left(\mu^{SE}\right)\right]=\mathbb{E}\left[H\left(\mu^{SE}\right)\right]-\varphi(\lambda)=\sum_{s,e}P\{s,e\}H\left(\mu^{se}\right)-\varphi(\lambda)$. For $s=+$ and $e=\hbar$, we have

$$\begin{aligned}\frac{d}{d\xi}\mathbb{P}\{s=+,e=\hbar\}H\left(\mu^{+\hbar}\right) &= \left(\theta\left(\left(1-\gamma^{\hbar}\right)\mu+\gamma^{\hbar}(1-\mu)\right)-(1-\theta)\frac{1}{2}\right)H\left(\mu^{+\hbar}\right)\\ &\quad+\frac{\frac{1}{2}\theta\mu(1-\theta)(1-\mu)\left(1-2\gamma^{\hbar}\right)}{\theta\xi\left(\left(1-\gamma^{\hbar}\right)\mu+\gamma^{\hbar}(1-\mu)\right)+(1-\theta)\frac{1}{2}(1-\xi)}\ln\left(\frac{1-\mu^{+\hbar}}{\mu^{+\hbar}}\right)\end{aligned}$$

Taking the second-order derivative, we obtain

$$\begin{aligned}&\left(\theta\left(\left(1-\gamma^{\hbar}\right)\mu+\gamma^{\hbar}\left(1-\mu\right)\right)-\left(1-\theta\right)\frac{1}{2}\right)\ln\left(\frac{1-\mu^{+\hbar}}{\mu^{+\hbar}}\right)\frac{d}{d\xi}\mu^{+\hbar}\\&+\ln\left(\frac{1-\mu^{+\hbar}}{\mu^{+\hbar}}\right)\frac{d}{d\xi}\frac{\frac{1}{2}\theta\mu\left(1-\theta\right)\left(1-\mu\right)\left(1-2\gamma^{\hbar}\right)}{\theta\xi\left(\left(1-\gamma^{\hbar}\right)\mu+\gamma^{\hbar}\left(1-\mu\right)\right)+\left(1-\theta\right)\frac{1}{2}\left(1-\xi\right)}\\&+\frac{\frac{1}{2}\theta\mu\left(1-\theta\right)\left(1-\mu\right)\left(1-2\gamma^{\hbar}\right)}{\theta\xi\left(\left(1-\gamma^{\hbar}\right)\mu+\gamma^{\hbar}\left(1-\mu\right)\right)+\left(1-\theta\right)\frac{1}{2}\left(1-\xi\right)}\frac{\mu^{+\hbar}}{1-\mu^{+\hbar}}\left(\frac{d}{d\xi}\frac{1-\mu^{+\hbar}}{\mu^{+\hbar}}\right)\end{aligned}$$

which gives

$$\begin{aligned}&\ln\left(\frac{1-\mu^{+\hbar}}{\mu^{+\hbar}}\right)\frac{\frac{1}{2}\theta\mu\left(1-\theta\right)\left(1-\mu\right)\left(1-2A\right)\left(\theta\left(\left(1-A\right)\mu+A\left(1-\mu\right)\right)-\left(1-\theta\right)\frac{1}{2}\right)}{\left(\theta\xi\left[\left(1-\gamma^{\hbar}\right)\mu+\gamma^{\hbar}\left(1-\mu\right)\right]+\left(1-\theta\right)\frac{1}{2}\left(1-\xi\right)\right)^{2}}\\&-\ln\left(\frac{1-\mu^{+\hbar}}{\mu^{+\hbar}}\right)\frac{\frac{1}{2}\theta\mu\left(1-\theta\right)\left(1-\mu\right)\left(1-2A\right)\left(\theta\left(\left(1-A\right)\mu+A\left(1-\mu\right)\right)-\left(1-\theta\right)\frac{1}{2}\right)}{\left(\theta\xi\left(\left(1-A\right)\mu+A\left(1-\mu\right)\right)+\left(1-\theta\right)\frac{1}{2}\left(1-\xi\right)\right)^{2}}\\&-\frac{\left(\frac{1}{2}\theta\mu\left(1-\theta\right)\left(1-\mu\right)\left(1-2A\right)\right)^{2}}{\left(\theta\xi\left[\left(1-\gamma^{\hbar}\right)\mu+\gamma^{\hbar}\left(1-\mu\right)\right]+\left(1-\theta\right)\frac{1}{2}\left(1-\xi\right)\right)^{3}}\frac{1}{\mu^{+\hbar}\left(1-\mu^{+\hbar}\right)}<0\end{aligned}$$

since the first two terms cancel each other. This means that the term $\mathbb{P}\left\{s=+,e=\hbar\right\}H\left(\mu^{+\hbar}\right)$ is concave when $\mu^{+\hbar}$ is within the search interval. The concavity of the other terms can be shown similarly. Since all terms are concave, $\mathbb{E}\left[H\left(\mu^{SE}\right)\right]$ and hence $C\left(\xi\right)$ is concave in ξ as $\varphi\left(\lambda\right)$ is constant. This is also the case when one or more of the posteriors go out of the search range as ξ increases, since their contribution to the expected cognitive effort will be zero. Therefore, $C\left(\xi\right)$ is piecewise concave in at most three regions.

We show that C is unimodal in the first interval. Observe that $\mathbb{E}\left[\mu^{SE}\right]=\mu$ for all $\xi<\underline{\xi}$. Since H is strictly concave, the behavior of $\mathbb{E}\left[H\left(\mu^{SE}\right)\right]$ with respect to ξ is completely determined by the variance of μ^{SE}. In particular, the higher the variance of μ^{SE}, the lower the expectation $\mathbb{E}\left[H\left(\mu^{SE}\right)\right]$ and hence $C\left(\xi\right)=\lambda\left(E\left[H\left(\mu^{SE}\right)\right]-\varphi\left(\lambda\right)\right)$. Since $C\left(\xi\right)$ is strictly concave, we show that at $\xi=1/2$, variance of the posteriors are in decreasing in ξ which shows that C is increasing at $\xi=1/2$. Variance of the posteriors can be written as

$$Var\left(\mu^{SE}\right)=\sum_{s,e\in\{+,-\}\times\{\hbar,\ell\}}\mathbb{P}\{s,e\}\left(\mu^{se}-\mu\right)^{2}.$$

Assume that $s=+$. After some algebra, the first-order derivative of the $\mathbb{P}\left\{s=+,e=\hbar\right\}\left(\mu^{+\hbar}-\mu\right)^{2}$ term reduces to

$$\begin{aligned}&\frac{\partial}{\partial\xi}\left(\theta\xi\left(\left(1-\gamma^{\hbar}\right)\mu+\gamma^{\hbar}\left(1-\mu\right)\right)+\left(1-\theta\right)\frac{1}{2}\left(1-\xi\right)\right)\\&\qquad\left(\frac{\theta\xi\left(1-\gamma^{\hbar}\right)\mu+\left(1-\theta\right)\frac{1}{2}\left(1-\xi\right)\mu}{\theta\xi\left(\left(1-\gamma^{\hbar}\right)\mu+\gamma^{\hbar}\left(1-\mu\right)\right)+\left(1-\theta\right)\frac{1}{2}\left(1-\xi\right)}-\mu\right)^{2}\\&=\frac{\xi\left(\theta\mu\left(1-\mu\right)\left(1-2\gamma^{\hbar}\right)\right)^{2}}{\left(\theta\xi\left(\left(1-\gamma^{\hbar}\right)\mu+\gamma^{\hbar}\left(1-\mu\right)\right)+\left(1-\theta\right)\frac{1}{2}\left(1-\xi\right)\right)^{2}}\\&\qquad\left(\left(\theta\left(\left(1-\gamma^{\hbar}\right)\mu+\gamma^{\hbar}\left(1-\mu\right)\right)-\left(1-\theta\right)\frac{1}{2}\right)\xi+1-\theta\right)\end{aligned}$$

Similarly, the first order derivative of $\mathbb{P}\left\{s=+,e=\ell\right\}\left(\mu^{+\ell}-\mu\right)^{2}$ is

$$\frac{\partial}{\partial\xi}\left(\theta\left(1-\xi\right)\left(\left(1-\gamma^{\ell}\right)\mu+\gamma^{\ell}\left(1-\mu\right)\right)+\left(1-\theta\right)\frac{1}{2}\xi\right)$$

$$\left(\frac{\theta\left(1-\xi\right)\left(1-\gamma^{\ell}\right)\mu+\left(1-\theta\right)\frac{1}{2}\xi\mu}{\theta\left(1-\xi\right)\left(\left(1-\gamma^{\ell}\right)\mu+\gamma^{\ell}\left(1-\mu\right)\right)+\left(1-\theta\right)\frac{1}{2}\xi}-\mu\right)^{2}$$
$$=\left(\frac{\left(1-\xi\right)\left(\theta\mu\left(1-\mu\right)\left(1-2\gamma^{\ell}\right)\right)^{2}}{\left(\theta\left(1-\xi\right)\left(\left(1-\gamma^{\ell}\right)\mu+\gamma^{\ell}\left(1-\mu\right)\right)+\left(1-\theta\right)\frac{1}{2}\xi\right)^{2}}\right)$$
$$\left(\left(-\theta\left(\left(1-\gamma^{\ell}\right)\mu+\gamma^{\ell}\left(1-\mu\right)\right)+\left(1-\theta\right)\frac{1}{2}\right)\left(1-\xi\right)-\left(1-\theta\right)\right).$$

Evaluating these at $\xi=1/2$, and summing up, we obtain

$$\frac{\left(\theta\mu\left(1-\mu\right)\left(1-2\gamma^{\hbar}\right)\right)^{2}}{\left(\theta\left(\left(1-\gamma^{\hbar}\right)\mu+\gamma^{\hbar}\left(1-\mu\right)\right)+\left(1-\theta\right)\frac{1}{2}\right)^{2}}\left(\left(\theta\left(\left(1-\gamma^{\hbar}\right)\mu+\gamma^{\hbar}\left(1-\mu\right)\right)+\left(1-\theta\right)\frac{1}{2}\right)+\left(1-\theta\right)\right)$$
$$-\left(\frac{\left(\theta\mu\left(1-\mu\right)\left(1-2\gamma^{\ell}\right)\right)^{2}}{\left(\theta\left(\left(1-\gamma^{\ell}\right)\mu+\gamma^{\ell}\left(1-\mu\right)\right)+\left(1-\theta\right)\frac{1}{2}\right)^{2}}\right)\left(\left(\theta\left(\left(1-\gamma^{\ell}\right)\mu+\gamma^{\ell}\left(1-\mu\right)\right)-\left(1-\theta\right)\frac{1}{2}\right)+\left(1-\theta\right)\right)$$
$$<0$$

which is negative since $\gamma^{\hbar}>\gamma^{\ell}$. One can show the same result for the last two terms for $s=-$ using the same approach. This shows that the variance of posteriors decreases in ξ at $\xi=1/2$. Consequently, $C\left(\xi\right)$ is increasing in ξ. Since $C\left(\xi\right)$ is strictly concave for $\xi<\underline{\xi}$, there must be a threshold $\widehat{\xi}$ (possibly equal to $\underline{\xi}$) such that cognitive effort is increasing in ξ for $\xi<\widehat{\xi}$. Assume for instance that $\underline{\xi}=1$. Then all posteriors are in the search range for all explainability values ξ. At the other extreme, $\xi=1$, note that the derivative of the first term vanishes, and since the derivative of the second term is positive, the sum is positive. This is also valid for the variance terms for $s=-$. This means that the variance of posteriors is increasing in ξ, and consequently, $C\left(\xi\right)$ is decreasing in ξ at $\xi=1$. Since C is strictly concave in this case in $\xi\in\left(1/2,1\right)$, C is strictly unimodal. Therefore, there always exists a non-empty interval for ξ such that $\Delta C\left(\xi\right)$ is positive.

Proof of Proposition 7 Given $s=+,\mu^{+\hbar}$ increases and $\mu^{+\ell}$ decreases in ξ per Lemma 1. Furthermore, $\mathbb{E}_{E}\left[\mu^{SE}\right]=\mu^{S}$ for all $\xi\geq 1/2$. This means that although the means of posteriors $\mu^{+\hbar}$ and $\mu^{+\ell}$ stay constant, their variability increases as they move away from each other. Since H is strictly concave, $\mathbb{E}_{E}\left[H\left(\mu^{+E}|S=+\right)\right]$ decreases in ξ. Same argument is valid for $s=-$ and hence the expectation $\lambda_{e}\mathbb{E}_{S}\left[\left(H\left(\mu^{S}\right)-\mathbb{E}_{E}\left[H\left(\mu^{SE}\right)\right]\right)\right]$ increases in ξ. Since $\mu^{SE}=\mu^{S}$ at $\xi=1/2$, $C_{e}\left(1/2\right)=0$.

LEMMA A.2. *$A(\xi)-C(\xi)$ is increasing in ξ.*

Proof of Lemma A.2 Let $V_{h}^{\star}\left(\mu\right)=A_{h}^{\star}\left(\mu\right)-C_{h}^{\star}\left(\mu\right)$ and $\overline{V}\left(\xi\right)=A\left(\xi\right)-C\left(\xi\right)$. We first show that $V_{h}^{\star}$ is convex in $[0,1]$. Assume that $\mu\leq\underline{\mu}$. Then, $V_{h}^{\star}\left(\mu\right)=1-\mu$, which is linearly decreasing in μ. Assume that $\mu\in\left(\underline{\mu},\overline{\mu}\right)$. Then $A_{h}^{\star}\left(\mu\right)=\frac{e^{1/\lambda}}{e^{1/\lambda}+1}$ which is constant in μ, and since $C_{h}^{\star}\left(\mu\right)=\lambda\left(H\left(\mu\right)-\varphi\left(\lambda\right)\right)$ is strictly concave (due to H being strictly concave), $V_{h}^{\star}\left(\mu\right)$ is strictly convex. Finally, when $\mu\geq\overline{\mu}$, $V_{h}^{\star}\left(\mu\right)=\mu$ which is linearly increasing. To show convexity in $[0,1]$, we should check the derivative of V_{h} at $\underline{\mu}$ and $\overline{\mu}$. Assume $\mu\in\left(\underline{\mu},\overline{\mu}\right)$. Then the first order derivative of $V_{h}^{\star}$ is $-\lambda H'\left(\mu\right)=-\lambda\ln\left(\left(1-\mu\right)/\mu\right)$. At, $\mu=\underline{\mu}$, the first-order derivative becomes

$$-\lambda\ln\frac{1-\frac{1}{e^{1/\lambda}+1}}{\frac{1}{e^{1/\lambda}+1}}=-\lambda\ln e^{\frac{1}{\lambda}}=-1$$

which is the same as the derivative of $1-\mu$. Since $V_h^\star(\mu)$ is strictly convex in $\left(\underline{\mu},\overline{\mu}\right)$, the derivative will increase until $\mu=\overline{\mu}$ where it will be equal to 1, which is also the derivative of μ (i.e., $V_h^\star(\mu)$ when $\mu\geq\overline{\mu}$). Therefore $V_h^\star(\mu)$ is convex in $[0,1]$. We write the expected value as

$$\overline{V}(\xi)=\mathbb{E}\left[V_h^\star\left(\mu^{SE}\right)\right]=\mathbb{P}\{S=+\}\,\mathbb{E}\left[V_h^\star\left(\mu^{+E}\right)|S=+\right]+\mathbb{P}\{S=-\}\,\mathbb{E}\left[V_h^\star\left(\mu^{-E}\right)|S=-\right].$$

Note that $\mu^{+\hbar}$ increases and $\mu^{+\ell}$ decreases in ξ per Lemma 1. In other words, as ξ increases, posteriors move away from their mean $\mathbb{E}[\mu^{SE}|S=+]=\mu^+$ which is constant in ξ. Since the variability of mean-preserving posteriors $\mu^{+\hbar}$ and $\mu^{+\ell}$ increases and $V_h^\star$ is convex, the expectation $\mathbb{E}[V_h^\star(\mu^{+E})|S=+]$ increases. This also applies to $\mathbb{E}[V_h^\star(\mu^{-E})|S=-]$. Since the weights $\mathbb{P}\{S=s\}$ are constant in ξ, $V(\xi)$ increases in ξ.

Proof of Proposition 8 $V_e(\xi)$ is decreasing in λ_e and $V_e(1/2)$ is constant. Assume that $\lambda_e=\lambda$, $\mu>1/2$ and $\xi<\underline{\xi}$. Then,

$$\begin{aligned}C_e(\xi)+C(\xi)&=\lambda\left(\mathbb{E}\left[H\left(\mu^{S}\right)\right]-\mathbb{E}\left[H\left(\mu^{SE}\right)\right]+\mathbb{E}\left[H\left(\mu^{SE}\right)\right]-\varphi(\lambda)\right)\\&=\lambda\left(\mathbb{E}\left[H\left(\mu^{S}\right)\right]-\varphi(\lambda)\right)=C(1/2)\end{aligned}$$

which is constant in ξ. Note that $A(\xi)=A(1/2)=e^{1/\lambda}/\left(e^{1/\lambda}+1\right)$. Therefore, $V_e(\xi)=V(1/2)$ and hence $\Delta V_e(\xi)=0$. Assume now that $\underline{\xi}\leq\xi<\overline{\xi}$. Then, $\Delta V_e(\xi)$ can be written as

$$\begin{aligned}\Delta V_e(\xi)&=A(\xi)-A(1/2)-C(\xi)+C(1/2)-C_e(\xi)\\&=\mathbb{P}\{+,\hbar\}\left(\mu^{+\hbar}-\frac{e^{1/\lambda}}{e^{1/\lambda}+1}+H\left(\mu^{+\hbar}\right)-\varphi(\lambda)\right).\end{aligned}$$

The first term is positive. We show that the second term is negative. Taking the first-order derivative of the second term with respect to ξ, we obtain

$$\begin{aligned}&\frac{\partial}{\partial\xi}\mu^{+\hbar}+\frac{\partial}{\partial\xi}H\left(\mu^{+\hbar}\right)\\&=\left(1+H'\left(\mu^{+\hbar}\right)\right)\frac{\partial}{\partial\xi}\mu^{+\hbar}=\left(1+\ln\left(\frac{1-\mu^{+\hbar}}{\mu^{+\hbar}}\right)\right)\frac{\partial}{\partial\xi}\mu^{+\hbar}\end{aligned}$$

The second term is positive per Lemma 1. As ξ increases, $\mu^{+\hbar}$ increases, and hence the log term is negative and decreasing. The highest possible value for the log term is $\ln\left((1-\overline{\mu})/\overline{\mu}\right)$ since we assumed $\underline{\xi}\leq\xi<\overline{\xi}$. Plugging in $\overline{\mu}=e^{1/\lambda}/\left(e^{1/\lambda}+1\right)$, we obtain $\ln\left(\left(1-\mu^{+\hbar}\right)/\mu^{+\hbar}\right)=-1/\lambda$. Note that $\lambda=1$ violates our complementarity assumptions. To see this, note that

$$\begin{aligned}\mu^+=\frac{\mu\theta+\frac{1}{2}(1-\theta)\mu}{\mu\theta+\frac{1}{2}(1-\theta)}&<\overline{\mu}=\frac{e^{1/\lambda}}{e^{1/\lambda}+1}\\\left(\mu\theta+\frac{1}{2}(1-\theta)\mu\right)\left(e^{1/\lambda}+1\right)&<e^{1/\lambda}\left(\mu\theta+\frac{1}{2}(1-\theta)\right)\\\frac{\mu(1+\theta)}{(1-\mu)(1-\theta)}&<e^{1/\lambda}.\end{aligned}$$

At $\lambda=1$, μ can be at most $e/(e+1)$. Plugging this in, we have

$$\frac{\frac{e}{e+1}(1+\theta)}{\left(1-\frac{e}{e+1}\right)(1-\theta)}<e\Leftrightarrow\frac{1+\theta}{1-\theta}<1$$

which is not correct. Therefore, λ should be less than 1, $1-1/\lambda<0$ and hence $\Delta V_e(\xi)<0$. One can show that this holds for $\xi\geq\overline{\xi}$ as well.

Note that at $\lambda_e=0$, $\Delta V_e(\xi)=A(\xi)-C(\xi)-A(1/2)+C(1/2)>0$ since the expected value $A(\xi)-C(\xi)$ is increasing in ξ Lemma A.2. At $\lambda_e=\lambda$, we showed that $\Delta V_e(\xi)\leq 0$. Since $V_e(\xi)$ is decreasing in λ_e, there must exists a threshold $\lambda_e^*(\xi)$ on λ_e such that $\Delta V_e(\xi)=0$. Equating $V_e(\xi)$ and $V_e(1/2)$, we have

$$A(\xi)-C(\xi)-C_e(\xi)=A(1/2)-C(1/2)-C_e(1/2)$$
$$A(\xi)-C(\xi)-\lambda_e\mathbb{E}_S\left[\left(H\left(\mu^S\right)-\mathbb{E}_E\left[H\left(\mu^{SE}\right)\right]\right)\right]=A(1/2)-C(1/2)$$

which gives

$$\lambda_e^*(\xi)=\frac{A(\xi)-C(\xi)-(A(1/2)-C(1/2))}{\mathbb{E}_S\left[\left(H\left(\mu^S\right)-\mathbb{E}_E\left[H\left(\mu^{SE}\right)\right]\right)\right]}.$$

Note that the numerator is always positive as $A(\xi)-C(\xi)$ increases in ξ per Lemma A.2 and $\mathbb{E}_S\left[\left(H\left(\mu^S\right)-\mathbb{E}_E\left[H\left(\mu^{SE}\right)\right]\right)\right]>0$ due to the concavity of H, μ^{SE} being mean-preserving and having increasing variance in ξ.

To find how this threshold changes with respect to ξ, we use the implicit function theorem. We have

$$\frac{\partial}{\partial\xi}\Delta V_e(\xi)_{|\lambda_e=\lambda_e^*(\xi)}\frac{\partial\lambda_e^*(\xi)}{\partial\xi}+\frac{\partial}{\partial\lambda_e}\Delta V_e(\xi)_{|\lambda_e=\lambda_e^*(\xi)}=0$$

which gives

$$\frac{\partial\lambda_e^*(\xi)}{\partial\xi}=-\frac{\frac{\partial}{\partial\lambda_e}\Delta V_e(\xi)_{|\lambda_e=\lambda_e^*(\xi)}}{\frac{\partial}{\partial\xi}\Delta V_e(\xi)_{|\lambda_e=\lambda_e^*(\xi)}}.$$

Note that

$$\frac{\partial}{\partial\lambda_e}\Delta V_e=-\mathbb{E}_S\left[\left(H\left(\mu^S\right)-\mathbb{E}_E\left[H\left(\mu^{SE}\right)\right]\right)\right]<0$$

since $\mathbb{E}_S\left[\left(H\left(\mu^S\right)-\mathbb{E}_E\left[H\left(\mu^{SE}\right)\right]\right)\right]>0$ due to Jensen's inequality as H is strictly concave. Furthermore,

$$\frac{\partial}{\partial\xi}\Delta V_e=\frac{\partial}{\partial\xi}\left(A(\xi,\lambda_e^*)-C(\xi,\lambda_e^*)-C_e(\xi,\lambda_e^*)\right)\leq 0$$

since at $\xi=1/2$, ΔV is increasing as $V_e(\xi)$ is increasing and $C_e(\xi)=0$. Consequently when $\Delta V=0$, the derivative is nonincreasing. Therefore, $\frac{\partial\xi^*(\lambda_e)}{\partial\lambda_e}<0$.

Proof of Proposition 9 Per Proposition 8, when $\lambda_e<\lambda_e^*$, $\Delta V_e>0$ and hence $O_e(\xi)=O(\xi)$, $U_e(\xi)=U(\xi)$ and $A_e(\xi)=A(\xi)$ as the DM uses machine's explanation but the extra effort of eliciting it does not affect the DM's reliance behavior (as overreliance, underreliance, and accuracy do not depend on λ_e). When $\lambda_e\geq\lambda_e^*$, the DM does not elicit machine's explanation, and hence $O_e(\xi)=O(1/2)$, $U_e(\xi)=U(1/2)$ and $A_e(\xi)=A(1/2)$. As λ_e^* is decreasing on ξ per Proposition 8, there exists a threshold ξ^* altering the DM's decision on eliciting the machine's explanation. To show how this threshold changes with λ_e, assume that $\xi<\underline{\xi}$. Then,

$$\begin{aligned}V_e(\xi)&=\frac{e^{1/\lambda}}{e^{1/\lambda}+1}+\lambda\varphi(\lambda)-\lambda\mathbb{E}\left[H\left(\mu^{se}\right)\right]-\lambda_e\mathbb{E}\left[H\left(\mu^{s}\right)-H\left(\mu^{se}\right)\right]\\&=\frac{e^{1/\lambda}}{e^{1/\lambda}+1}+\lambda\varphi(\lambda)-\lambda_e\mathbb{E}\left[H\left(\mu^{s}\right)\right]-(\lambda+\lambda_e)\,E\left[\mathbb{E}\left(\mu^{se}\right)\right].\end{aligned}$$

The only term that depends on ξ is $\mathbb{E}\left[H\left(\mu^{se}\right)\right]$ which is decreasing in ξ as $\mathbb{E}\left[\mu^{se}\right]=\mu$, the variability of μ^{se} increases in ξ and H is a concave function. Therefore, $V_e(\xi)$ is increasing in ξ when $\xi<\underline{\xi}$. Using the implicit function theorem, we have

$$\frac{\partial}{\partial\xi}\Delta V\left(\xi^*,\lambda_e\right)\frac{\partial\xi^*}{\partial\lambda_e}+\frac{\partial}{\partial\lambda_e}\Delta V\left(\xi^*,\lambda_e\right)=0$$

which gives

$$\frac{\partial \xi^*}{\partial \lambda_e} = -\frac{\frac{\partial}{\partial \lambda_e}\Delta V\left(\xi^*, \lambda_e\right)}{\frac{\partial}{\partial \xi}\Delta V\left(\xi^*, \lambda_e\right)}.$$

Note that

$$\frac{\partial}{\partial \lambda_e}\Delta V = -\mathbb{E}_S\left[\left(H\left(\mu^S\right) - \mathbb{E}_E\left[H\left(\mu^{SE}\right)\right]\right)\right] < 0$$

since $\mathbb{E}_S\left[\left(H\left(\mu^S\right) - \mathbb{E}_E\left[H\left(\mu^{SE}\right)\right]\right)\right] > 0$ due to Jensen's inequality as H is strictly concave. Furthermore,

$$\frac{\partial}{\partial \xi}\Delta V = \frac{\partial}{\partial \xi}\left(A\left(\xi^*\right) - C\left(\xi^*\right) - C_e\left(\xi^*\right)\right) \leq 0$$

since at $\xi = 1/2$, ΔV_e is increasing as $V_e\left(\xi\right)$ is increasing and $C_e\left(\xi\right) = 0$. Consequently when $\Delta V_e = 0$, the derivative is nonincreasing. Therefore, $\frac{\partial \xi^*\left(\lambda_e\right)}{\partial \lambda_e} < 0$.

Proof of Proposition 10 Assume that $\xi < \xi^*_{\lambda_e}$. Then $\Delta C_t\left(\xi\right) = C_e\left(\xi\right) + C\left(\xi\right) - C\left(1/2\right) = C_e\left(\xi\right) + \Delta C\left(\xi\right)$ since $C_e\left(1/2\right) = 0$. Note that Theorem 3 continues to hold as $C_e\left(\xi\right) > 0$ for any $\xi > 1/2$. Therefore, the existence of an interval $I_e \subset (1/2, 1]$ for explainability ξ is guaranteed within which $\Delta C_t\left(\xi\right) > 0$.

Proof of Proposition 11 Per Proposition 2, we assume that for $\mu = 1/2$, $\mu^+ < \overline{\mu}$ and $\mu^- > \underline{\mu}$ as both the thresholds and the posteriors for $\xi = 1/2$ are symmetric around $\mu = 1/2$. Note that $\mu^- < \underline{\mu}$ for $\mu = \underline{\mu} < 1/2$. Since μ^- is decreasing in μ, there must be an interval for μ within $\left(\underline{\mu}, 1/2\right)$ such that $\mu^- < \underline{\mu}$ and $\mu^+ \in \left(\underline{\mu}.\overline{\mu}\right)$. Assume now that $\mu < 1/2$. Then note that $\underline{\mu}$ decreases and $\underline{\mu}$ increases in λ. Since they are symmetric around $\mu = 1/2$, for $\mu < 1/2$, as λ increases, the lower threshold $\underline{\mu}$ will intersect μ^- first. Therefore, there must exist a threshold $\overline{\lambda}_s$ (which is higher than the complementarity threshold $\overline{\lambda}$ as we violate complementarity), such that $\mu^- < \underline{\mu}$ and $\mu^+ \in \left(\underline{\mu}.\overline{\mu}\right)$ for $\overline{\lambda} < \lambda < \overline{\lambda}_s$.

Proof of Proposition 12 When $\mu^- = \mu^{-e} < \underline{\mu}$ for $\xi \in [1/2, 1]$ and $e \in \{\hbar, \ell\}$, overreliance is $O\left(\xi\right) = \mathbb{E}\left[\alpha^\star_h\left(\mu^{+E}\right)\mathbb{I}_{\{S=+\}} + \beta^\star_h\left(\mu^-\right)\mathbb{I}_{\{S=-\}}\right]$ where only the first term inside the expectation varies with ξ. This means that as explainability ξ increases, only $\mu^{+\hbar}$ will intersect the belief threshold $\overline{\mu}$ as $\mu^{+\hbar}$ converges to 1 as ξ approaches 1. As before, we label $\underline{\xi}$ such that $\mu^{+\hbar} = \overline{\mu}$. Note that for $\xi \leq \underline{\xi}$, both $\mu^{+\hbar}$ and $\mu^{+\ell}$ are within the search interval and since $\alpha^\star_h$ is linear, $\mathbb{E}\left[\alpha^\star_h\left(\mu^{+E}\right)\mathbb{I}_{\{S=+\}}\right] = \alpha^\star_h\left(\mathbb{E}\left[\mu^{+E}\mathbb{I}_{\{S=+\}}\right]\right) = \alpha^\star_h\left(\mu^+\right)$ which is constant in ξ. For $\xi > \underline{\xi}$, $\mu^{+\hbar} > \overline{\mu}$ and $\alpha^\star_h\left(\mu^{+\hbar}\right) = 1 - \mu^{+\hbar}$. Since $\mu^{+\hbar}$ increases in ξ, $\alpha^\star_h\left(\mu^{+\hbar}\right)$ decreases in ξ. Similarly, since $\mu^{+\ell}$ decreases in ξ and $\alpha^\star_h\left(\mu^{+\hbar}\right)$ increases in $\mu^{+\hbar}$, $\alpha^\star_h\left(\mu^{+\hbar}\right)$ decreases in ξ. Since both $\alpha^\star_h\left(\mu^{+\hbar}\right)$ and $\alpha^\star_h\left(\mu^{+\ell}\right)$ decrease in ξ, their expectation also decreases. Hence, $O\left(\xi\right)$ decreases in ξ. Since $\beta^\star_h$ and $A^\star_h$ are linear in belief within the search region, $U\left(\xi\right)$ and $A\left(\xi\right)$ are also constant in ξ when $\xi \leq \underline{\xi}$. However, the affects are reversed for $\xi > \underline{\xi}$. For $\mu^{+\hbar} > \overline{\mu}$, underreliance term $\mathbb{P}\{+, \hbar\}\,\beta^\star_h\left(\mu^{+\hbar}\right) = 0$ and $\mathbb{P}\{+, \hbar\}\,A^\star_h\left(\mu^{+\hbar}\right)$ increases in $\mu^{+\hbar}$. On the other hand, $\mathbb{P}\{+, \ell\}\,\beta^\star_h\left(\mu^{+\ell}\right)$ decreases and $\mathbb{P}\{+, \ell\}\,A^\star_h\left(\mu^{+\ell}\right)$ is constant in $\mu^{+\ell}$. Therefore, $U\left(\xi\right)$ and $A\left(\xi\right)$ are increasing in ξ when $\xi > \underline{\xi}$.

Proof of Proposition 13 Assume $\mu^{-e} = \mu^- < \underline{\mu}$ for $\xi \geq 1/2$, $\mu < 1/2$ and $\mu^+ \in \left(\underline{\mu}, \overline{\mu}\right)$. Then the DM's expected cognitive cost as a function of explainability ξ is $C\left(\xi\right) = \mathbb{P}\{S = +\}\,\mathbb{E}\left[C^\star_h\left(\mu^{+E}\right)|S = +\right]$ since $C^\star_h\left(\mu^-\right) = 0$. Let $\underline{\xi}$ be such that $\mu^{+\hbar} = \overline{\mu}$. Assume that $\xi < \underline{\xi}$ and hence both $\mu^{+\hbar}$ and $\mu^{+\ell}$ are in the search interval. Note that $\mu^{+\hbar}$ increases and $\mu^{+\ell}$ decreases in ξ and they move away from their mean μ^+ which is constant in ξ. Since their variability increases in ξ and $C^\star_h$ is strictly concave, $\mathbb{E}\left[C^\star_h\left(\mu^{+E}|S = +\right)\right]$ decreases in ξ. Assume now that λ is sufficiently high such that $\mu^{+\hbar} = \mu^{+\ell} = \overline{\mu}$ for $\xi = 1/2$. At this point $C\left(\xi\right) = 0$

and as ξ is increased, the cognitive effort increases (as C is nonnegative) and hence $\Delta C(\xi) > 0$. Since C is continuous in ξ and λ and piecewise concave in ξ, there must exist an interval I where the cognitive effort is positive for a sufficiently high information cost λ value.

Lemma A.3. *Given μ, λ and δ, we have*

$$A_h^\star(\mu) = \begin{cases} 1-\mu & \text{if } \mu \le \underline{\mu}_\delta \\ (1-\underline{\mu}_\delta)(1-p_\delta) + \overline{\mu}_\delta p_\delta & \text{if } \mu \in (\underline{\mu}_\delta, \overline{\mu}_\delta) \\ \mu & \text{if } \mu \ge \overline{\mu} \end{cases}$$

$$C_h^\star(\mu) = \begin{cases} \lambda(H(\mu) - \varphi_\delta(\mu,\lambda)) & \text{if } \mu \in (\underline{\mu}_\delta, \overline{\mu}_\delta) \\ 0 & \text{otherwise} \end{cases}$$

$$\alpha_h^\star = \begin{cases} 0 & \text{if } \mu \le \underline{\mu}_\delta \\ (1-\overline{\mu}_\delta)p_\delta & \text{if } \mu \in \left(\underline{\mu}_\delta, \overline{\mu}_\delta\right) \\ 1-\mu & \text{if } \mu \ge \overline{\mu}_\delta \end{cases} \qquad \beta_h^\star = \begin{cases} \mu & \text{if } \mu \le \underline{\mu}_\delta \\ \overline{\mu}_\delta(1-p_\delta) & \text{if } \mu \in \left(\underline{\mu}_\delta, \overline{\mu}_\delta\right) \\ 0 & \text{if } \mu \ge \overline{\mu}_\delta \end{cases}$$

where

$$p_\delta = \frac{\mu}{1-e^{-\delta/\lambda}} - \frac{1-\mu}{e^{1/\lambda}-1}$$

and

$$\varphi_\delta(\mu,\lambda) = p_\delta H(\overline{\mu}_\delta) - (1-p_\delta)H(\underline{\mu}_\delta)$$

and

$$\underline{\mu}_\delta = \frac{1-e^{-\delta/\lambda}}{e^{1/\lambda}-e^{-\delta/\lambda}} \qquad \text{and} \qquad \overline{\mu}_\delta = e^{1/\lambda}\frac{1-e^{-\delta/\lambda}}{e^{1/\lambda}-e^{-\delta/\lambda}}.$$

Proof of Lemma A.3 Using Theorem 1 in Matějka and McKay (2015), the DM's conditional probability of choosing $a=y$ given $\omega=1$ and $\omega=0$ can be written as

$$\mathbb{P}\{a=y|\omega=1\} = \frac{p_y e^{1/\lambda}}{p_y e^{1/\lambda}+1-p_y}$$
$$\mathbb{P}\{a=y|\omega=0\} = \frac{p_y}{p_y+(1-p_y)e^{\delta/\lambda}}$$

where $p_y = \mathbb{P}\{a=y\}$ is the DM's unconditional probability of choosing y. Then p_y should satisfy the consistency condition

$$p_y = (1-\mu)\frac{p_y}{p_y+(1-p_y)e^{\delta/\lambda}} + \mu\frac{p_y e^{1/\lambda}}{p_y e^{1/\lambda}+1-p_y}.$$

Solving this for p_y, we obtain

$$p_\delta = \frac{\mu}{1-e^{-\frac{\delta}{\lambda}}} - \frac{1-\mu}{e^{\frac{1}{\lambda}}-1}$$

which should be within the interval $[0,1]$. Then $p_\delta \le 0$ gives $\mu \le \underline{\mu}_\delta$ and $p_\delta \ge 1$ gives $\mu \ge \overline{\mu}_\delta$. Therefore, the optimal choice probability is characterized as

$$p_y^\star = \begin{cases} 0 & \text{if } \mu \le \underline{\mu}_\delta \\ p_\delta & \text{if } \mu \in \left(\underline{\mu}_\delta, \overline{\mu}_\delta\right) \\ 1 & \text{if } \mu \ge \overline{\mu}_\delta \end{cases}$$

The belief thresholds correspond to the DM's optimal posterior beliefs given their choice, such that $\overline{\mu}_\delta = \mathbb{P}\{\omega=1|a=y\}$ and $\underline{\mu}_\delta = \mathbb{P}\{\omega=1|a=n\}$. Then the DM's decision accuracy is $\mathbb{P}\{\omega=1|a=y\}\mathbb{P}\{a=y\} + \mathbb{P}\{\omega=0|a=n\}\mathbb{P}\{a=n\}$ which is $\overline{\mu}_\delta p_y^\star + \overline{\mu}_\delta\left(1-p_y^\star\right)$. In a similar manner, the DM's expected cognitive cost is $\lambda\left(H(\mu) - \left(p_y^\star H(\overline{\mu}_\delta) + \left(1-p_y^\star\right)H\left(\underline{\mu}_\delta\right)\right)\right)$ where the second term inside the parentheses is the DM's expected posterior entropy. Lastly, the DM's false-positive error is $\mathbb{P}\{\omega=0, a=y\} = (1-\overline{\mu}_\delta)p_y^\star$ and false-negative error is $\mathbb{P}\{\omega=1, a=n\} = \underline{\mu}_\delta\left(1-p_y^\star\right)$.

Proof of Proposition 14 Assume that $\xi < \underline{\xi}$. Then all posteriors μ^{se} are within the search interval. Since $\alpha_h^\star$ and $\beta_h^\star$ are linear in belief (as both $\overline{\mu}_\delta$ and $\underline{\mu}_\delta$ are constant and p_δ is linear in belief), overreliance becomes $O(\xi) = \alpha_h^\star \mathbb{E}[(\mu^{+E})\mathbb{I}_{\{S=+\}}] + \beta_h^\star \mathbb{E}[(\mu^{-E})\mathbb{I}_{\{S=-\}}] = \mathbb{E}[\alpha_h^\star(\mu^+)\mathbb{I}_{\{S=+\}} + \beta_h^\star(\mu^-)\mathbb{I}_{\{S=-\}}]$ which is constant in ξ. Same argument is valid for underreliance and accuracy. Assume now that $\underline{\xi} \le \xi < \overline{\xi}$ and $\mu^{+\hbar} \ge \overline{\mu}_\delta$. Then as before, $\mathbb{E}[\beta_h^\star(\mu^{-E})\mathbb{I}_{\{S=-\}}]$ is constant. On the other hand, $\mathbb{E}[\alpha_h^\star(\mu^{+E})\mathbb{I}_{\{S=+\}}] = \mathbb{P}\{+,\hbar\}\left(1-\mu^{+\hbar}\right) + \mathbb{P}\{+,\ell\}\alpha_h^\star\left(\mu^{+\ell}\right)$. As ξ increases, $1-\mu^{+\hbar}$ and $\alpha_h^\star\left(\mu^{+\ell}\right)$ decreases since $\mu^{+\hbar}$ increases and $\mu^{+\ell}$ decreases and $\alpha_h^\star$ is an increasing function. Therefore, overreliance decreases. This is reversed for underreliance as functions $\alpha_h^\star$ and $\beta_h^\star$ are switched. For accuracy, note that the first-order derivative of $\left(1-\underline{\mu}_\delta\right)(1-p_\delta) + \overline{\mu}_\delta p_\delta$ (i.e., accuracy when μ is in the search interval) with respect to μ can be found as

$$\frac{1-e^{\frac{1-\delta}{\lambda}}}{\left(1-e^{-\frac{\delta}{\lambda}}\right)\left(e^{\frac{1}{\lambda}}-1\right)} \tag{9}$$

Note that (9) is positive when $\delta > 1$, negative when $\delta < 1$ and 0 when $\delta = 1$ (our base case). For $\mu \ge \overline{\mu}_\delta$, accuracy becomes μ whose derivative is 1. Note that at $\mu = \overline{\mu}_\delta$, (9) is less than 1, for any $\delta > 0$. Therefore, the rate of increase in accuracy with respect to belief μ is always higher when $\mu \ge \overline{\mu}_\delta$. Since for $\xi < \underline{\xi}$, expected accuracy with respect to machine signals is constant, when $\xi \ge \underline{\xi}$, $\mu^{+\hbar}$ will be higher than $\overline{\mu}_\delta$ and the corresponding accuracy term starts to increase at a higher rate which also increases the expected accuracy. This is also the case for $\xi \ge \overline{\xi}$ where $\mu^{-\hbar}$ becomes less than the lower threshold $\underline{\mu}_\delta$. The corresponding accuracy term will then be $1-\mu^{-\hbar}$ which is also increasing in ξ. This means that the expected accuracy will increase at a higher rate with respect to ξ when $\xi \ge \underline{\xi}$. In a similar manner, when $\mu^{-\hbar} \le \underline{\mu}_\delta$, $\beta_h^\star\left(\mu^{-\hbar}\right) = \mu^{-\hbar}$ which is decreasing in ξ. Therefore, overreliance starts to decrease at a higher rate when $\xi \ge \underline{\xi}$. As before, the same argument is valid for underreliance, and it increases at a higher rate.

Proof of Proposition 15 Note that regarding the expected cognitive cost function, the only change from the base model is the expected posterior entropy terms given by $\varphi_\delta(\mu,\lambda) = p_\delta H(\overline{\mu}_\delta) + (1-p_\delta)H(\underline{\mu}_\delta)$ which now depends on μ for $\delta \ne 1$ linearly. Assume that all posteriors are in the search range. We already know that $\mathbb{E}[H(\mu^{SE})]$ is concave in ξ. As φ_δ is linear in μ, $\mathbb{E}[\varphi_\delta(\mu^{SE},\lambda)] = \varphi_\delta(\mathbb{E}[\mu^{SE}],\lambda) = \varphi_\delta(\mu,\lambda)$. Therefore, the cognitive effort function in the first interval is still concave for any $\delta > 0$. Furthermore, as in our base model, the derivative of $\lambda(\mathbb{E}[H(\mu^{SE})] - \varphi_\delta(\mu,\lambda))$ at $\xi = 1/2$ is equal to 0 since the derivative of the first term is zero (see the proof of Theorem 2), and the second term is constant. Furthermore, as in the base model, there exists a threshold $\overline{\lambda}_\delta$ on information cost (one of the new complementarity conditions) beyond which one of the posteriors when $\xi = 1/2$ fall out of the search range $(\underline{\mu}_\delta,\ \overline{\mu}_\delta)$. The reason is that as λ increases, posteriors μ^{se} remain constant while $\underline{\mu}_\delta$ increases and $\overline{\mu}_\delta$ decreases. Assume now that $\lambda \uparrow \overline{\lambda}_\delta$. Then the first threshold $\underline{\xi}$ on explainability approaches to $1/2$ and the first interval $\left[1/2, \underline{\xi}\right]$ vanishes. Assume that $\mu^{+\hbar}$ crosses the upper threshold $\overline{\mu}_\delta$ first as $\xi \uparrow \underline{\xi}$. Taking the first order derivative of the corresponding term, we have

$$\begin{aligned}
&\frac{\partial}{\partial\xi}\mathbb{P}\{+,\hbar\}\left(H\left(\mu^{+\hbar}\right) - \varphi_\delta\left(\mu^{+\hbar},\lambda\right)\right) \\
&= \left(H\left(\mu^{+\hbar}\right) - \varphi_\delta\left(\mu^{+\hbar},\lambda\right)\right)\frac{\partial}{\partial\xi}\mathbb{P}\{+,\hbar\} + \mathbb{P}\{+,\hbar\}\frac{\partial}{\partial\xi}\left(H\left(\mu^{+\hbar}\right) - \varphi_\delta\left(\mu^{+\hbar},\lambda\right)\right) \\
&= \mathbb{P}\{+,\hbar\}\frac{\partial}{\partial\xi}\left(H\left(\mu^{+\hbar}\right) - \varphi_\delta\left(\mu^{+\hbar},\lambda\right)\right)
\end{aligned} \tag{10}$$

as the first term vanishes since $H\left(\mu^{+\hbar}\right) \uparrow \varphi_\delta\left(\mu^{+\hbar}, \lambda\right)$ as $\xi \uparrow \underline{\xi}$ and hence $\mu^{+\hbar} \uparrow \overline{\mu}_\delta$. Note also that the cognitive effort $H\left(\mu^{+\hbar}\right) - \varphi_\delta\left(\mu^{+\hbar}, \lambda\right)$ associated with $\mu^{+\hbar}$ decreases as ξ increases and $\mu^{+\hbar}$ converges to $\overline{\mu}_\delta$. Therefore, (10) is negative. Since at $\xi = 1/2$, the first order derivative of the cognitive effort function is 0 and it is concave, $\xi = 1/2$ is the maximizer in the first interval. Therefore, it must be that the sum of the other terms $\mathbb{P}\{s, e\}\left(H\left(\mu^{se}\right) - \varphi_\delta\left(\mu^{se}, \lambda\right)\right)$ at $\xi = 1/2$ is positive. Therefore, the cognitive effort is increasing as $\xi \downarrow 1/2$ and $\lambda \uparrow \overline{\lambda}_\delta$ as in our base model. Due to the continuity of $C(\xi)$ in ξ and λ, there must exist a threshold on information cost λ beyond which there exists an interval on explainability such that the cognitive effort is higher relative to the no-explanation benchmark (i.e., $\Delta C > 0$).

Lemma A.4. *$O_h > U_h > 0$ for $\mu \neq 0.5$ and $O_h = U_h > 0$ for $\mu = 0.5$*

Proof of Lemma A.4

By the complementarity assumption, we have $\mu \in \left(\underline{\mu}, \overline{\mu}\right)$. Assume that $\mu > 0.5$. Then, per Proposition 1, we have $\alpha_h^\star(\mu) > \beta_h^\star(\mu)$. Furthermore, $P\{S = +\} = \mu\theta + \frac{1}{2}(1-\theta) > (1-\mu)\theta + \frac{1}{2}(1-\theta) = P\{S = -\}$. Then note that

$$\left(\alpha_h^\star(\mu) - \beta_h^\star(\mu)\right) P\{S = +\} > \left(\alpha_h^\star(\mu) - \beta_h^\star(\mu)\right) P\{S = -\}$$

which implies that

$$\begin{aligned} O_h &= P\{S = +\}\,\alpha_h^\star(\mu) + P\{S = -\}\,\beta_h^\star(\mu) \\ &> P\{S = +\}\,\beta_h^\star(\mu) + P\{S = -\}\,\alpha_h^\star(\mu)\,P\{S = -\} = U_h. \end{aligned}$$

The same result holds for $\mu < 0.5$ as $\alpha_h^\star(\mu) < \beta_h^\star(\mu)$ and $P\{S = +\} < P\{S = -\}$. Since $\alpha_h^\star(\mu) = \beta_h^\star(\mu)$ and $P\{S = +\} = P\{S = -\}$ for $\mu = 0.5$, we have $O_h = U_h$.

Proof of Corollary B.1

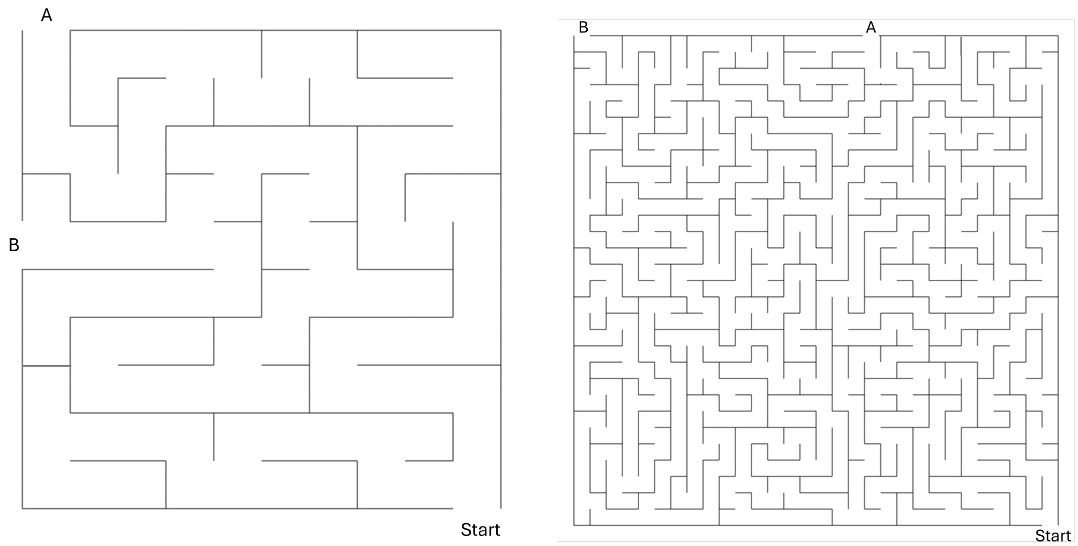
Per Corollary 1, $O(\xi) > U(\xi)$ for $\xi \in [0.5, 1)$ with $O(1) = U(1)$, and per Theorem 1, $O(\xi)$ is nonincreasing and $U(\xi)$ is nondecreasing. Since $O_h > U_h$ per Lemma A.4 and both O_h and U_h are constant in ξ, there must exist a threshold for ξ such that when ξ is higher than it, $O(\xi) - O_h > U(\xi) - U_h$.

Appendix B: Model Clarifications and Further Analysis

B.1. An Illustration of the Decision Task

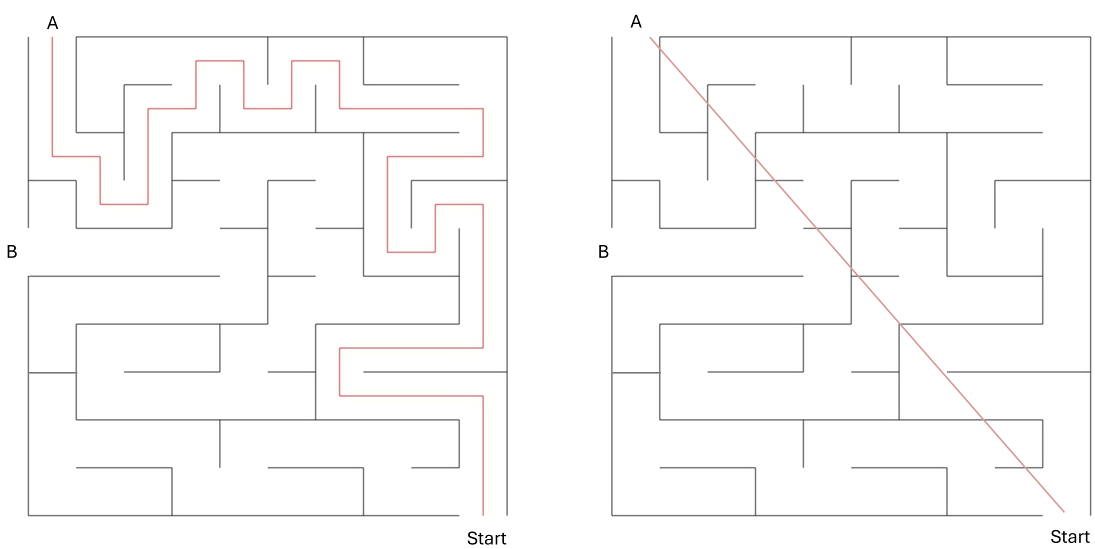
Figure B.1 is adapted from the maze experiments in Vasconcelos et al. (2023) where the DM aims to identify the correct exit towards which there is an uninterrupted path from the starting point. In this example, the two possible exits, labeled as A and B, form the action set $a \in \{y, n\}$, which is identical to the states of the world $\omega \in \{1, 0\}$. The DM does not have a priori information about the correct exit and thus $\mu = 1/2$. The more complex the maze, the more difficult it is to elicit information about possible paths, and hence the higher information cost λ. In this setup, the machine prediction is either A or B, which corresponds to $s = +$ and $s = -$, respectively.

Following Vasconcelos et al. (2023), the machine's explanation takes the form of visually highlighting a path to the machine's prediction A or B. Figure B.2 depicts a case where explainability is perfect ($\xi = 1$) and the machine indicates that A is the correct exit. A high-quality explanation ($e = \hbar$) perfectly signals



(a) Easy task (low λ) (b) Complex task (high λ)

Figure B.1 Illustration of a decision task with different complexity levels (Adapted from the maze experiments depicted in Vasconcelos et al. 2023)



(a) $e = \hbar$ (b) $e = \ell$

Figure B.2 Illustration of explanations with high explainability ($\xi \uparrow 1$) when the machine predicts exit A ($s = +$)

that the prediction is accurate by highlighting the correct path from the start to the exit (Figure B.2a). In contrast, a low-quality explanation ($e = \ell$) clearly indicates that the underlying rationale for the prediction is not sound, and hence that the machine is of low quality (Figure B.2b). It is worth noting that even though the machine is of low quality, the prediction can still be accurate. (In this example, the machine selects the exit with the longest Euclidean distance to the starting point; and since there is a priori no reason for the correct exit to be more distant to the starting point, the probability that the prediction is correct is 1/2, which exactly corresponds to machine's type $X = \mathfrak{l}$ in our setup). In this example, a non-perfect level of explainability ($1/2 < \xi < 1$) would be a partial path from the prediction back towards the start of the maze, but not quite reaching it.

Since the DM learns both the true state of the world ω and the machine's prediction quality X simultaneously, the DM needs to exert more effort to identify the correct gate when $e = \ell$, as it is relatively easier to infer that the prediction is poor ($X = \mathfrak{l}$). Conversely, when $e = \hbar$, more effort is needed to verify that the machine is accurate ($X = \mathfrak{h}$), though less effort may be needed to identify the correct gate. Nonetheless, as the information cost λ is uniform across these scenarios, our framework endogenously aggregates both types of effort without distinction.

In §6.2, we extend our model to a scenario where it is cognitively costly for the DM to understand the explanation signal E, where the information cost λ_e reflects the difficulty of processing explanations. For example, in the maze-solving task illustrated in Figure B.2a, while a low λ_e corresponds to highlighting the exit paths, a high λ_e would correspond written exit instructions such as "down, left, down, left, up, right, up, right, down, right, up, right, down, left, down, right, down, left, up, right, down, right, down, left, down", which are more cognitively demanding.

B.2. Incorporating Counterfactuals in Reliance Metrics

We define the counterfactual benchmarks for overreliance and underreliance as the probability of errors the DM would have made in the absence of the machine. Specifically, we define

$$O_h = \mathbb{E}\left[\alpha_h^\star(\mu)\mathbb{I}_{\{S=+\}} + \beta_h^\star(\mu)\mathbb{I}_{\{S=-\}}\right] \tag{11}$$

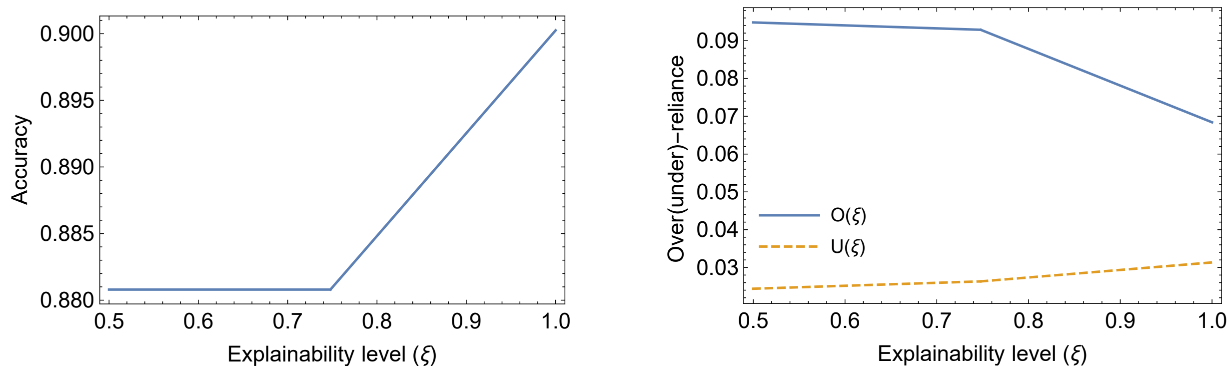
and

$$U_h = \mathbb{E}\left[\beta_h^\star(\mu)\mathbb{I}_{\{S=+\}} + \alpha_h^\star(\mu)\mathbb{I}_{\{S=-\}}\right]. \tag{12}$$

We then define the net overreliance and net underreliance as $O(\xi) - O_h$ and $U(\xi) - U_h$, respectively, capturing the additional errors attributable to the use of the machine. The following result is analogous to Corollary 1 and characterizes the relationship between net overreliance and net underreliance under these alternative specifications.

Corollary B.1. *There exists a threshold $\xi_c \in (0.5, 1]$ such that $O(\xi) - O_h > U(\xi) - U_h$ for $\xi < \xi_c$ and $O(\xi) - O_h \le U(\xi) - U_h$ otherwise.*

Corollary B.1 shows that, contrary to Corollary 1, net overreliance is not always higher than net overreliance, especially when the explainability level ξ is sufficiently high. This result is driven by the fact that the counterfactual benchmark for overreliance is higher than that for underreliance (see Lemma A.4 in Appendix A). Intuitively, when acting without the machine, the DM is more likely to make errors aligned with their prior belief—more false positives (resp. negatives) when $\mu > 0.5$ (resp. $\mu < 0.5$). Moreover, the machine would also be more likely to generate signals consistent with the DM's prior, leading to a higher counterfactual error rate for overreliance than for underreliance, as overreliance requires the DM to conform to the machine's prediction, while underreliance requires the DM to negate it.



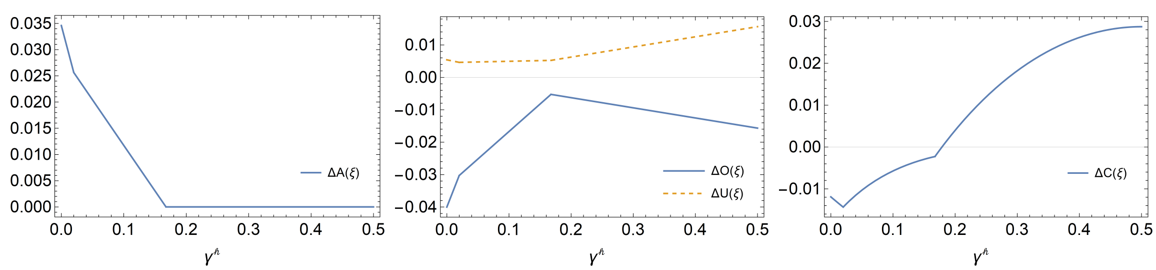
(a) Effect on accuracy (b) Effect on over(under)-reliance

Figure B.3 Effect of explainability ξ on over(under)-reliance and accuracy ($\mu = 0.65, \lambda = 0.5$, $\theta = 0.5$, $\gamma^{\hbar} = 0.1$, $\gamma^{\ell} = 0$).

B.3. Further Results for §6.1

Figure B.3 illustrates the impact of explainability ξ on the DM's reliance behavior when there is a negative explainability-accuracy tradeoff.

In this section, we further present a numerical analysis on the impact of the gap between $\gamma^{\hbar}$ and γ^{ℓ} on the DM's reliance behavior as well as cognitive effort (in §6.1). For simplicity, we assume that $\gamma^{\ell} = 0$ (such that a high-quality machine is accurate as in our baseline model when $e = \ell$) and solely vary $\gamma^{\hbar}$. As changing $\gamma^{\hbar}$ impacts the DM's reliance when $\xi = 1/2$, we consider how $\Delta O(\xi)$, $\Delta U(\xi)$, $\Delta A(\xi)$, and $\Delta C(\xi)$ change with respect to $\gamma^{\hbar}$ for a given ξ value. Figure B.4 illustrates these effects.



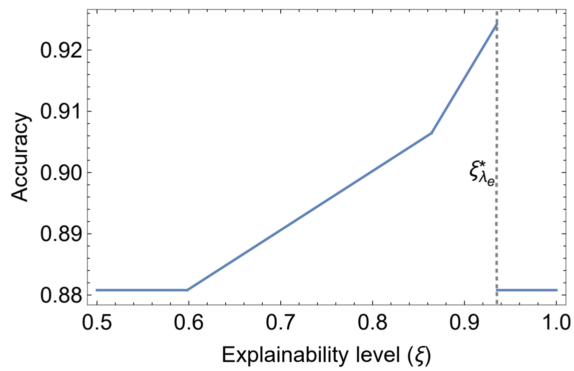
(a) Effect on accuracy (b) Effect on reliance (c) Effect on cognitive effort

Figure B.4 Effect of $\gamma^{\hbar}$ on reliance behavior and cognitive effort ($\mu = 0.65, \lambda = 0.5$, $\theta = 0.5$, $\xi = 0.9$).

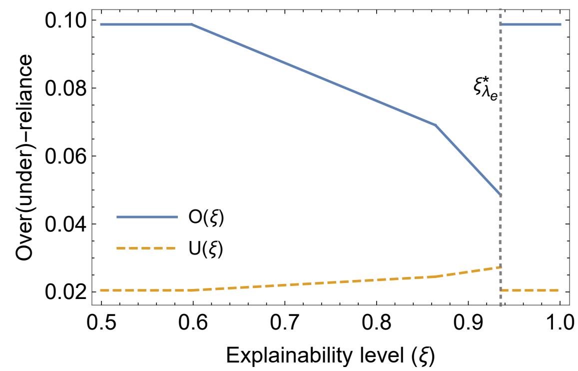
In particular, we observe that $\Delta O(\xi)$ is negative concave and piecewise linear in at most three regions in $\gamma^{\hbar}$. Furthermore, there exists a threshold $\overline{\gamma}^{\hbar}$ (possibly degenerate) such that $\Delta O(\xi)$ is increasing for $\gamma^{\hbar} < \overline{\gamma}^{\hbar}$ and decreases otherwise. This implies that the effects of explainability on reducing overreliance might be limited for moderate levels of $\gamma^{\hbar}$. Similarly, $\Delta U(\xi)$ is positive and convex in $\gamma^{\hbar}$ and is minimized for a moderate value of $\gamma^{\hbar}$. Accuracy, on the other hand, is convex decreasing and is equal to zero for sufficiently high values of $\gamma^{\hbar}$. This implies that as the negative correlation between explainability and accuracy increases (i.e., as $\gamma^{\hbar}$ increases), accuracy improvements brought about by explainability diminish. Finally, we observe that cognitive effort (relative to the no-explanation benchmark) is also non-monotonic, but increases in $\gamma^{\hbar}$ when $\gamma^{\hbar}$ is sufficiently high.

B.4. Further Results for §6.2

Figure B.5 below illustrates the impact of explainability on the DM's reliance behavior when processing machine explanations are costly, as outlined in §6.2.



(a) Effect on accuracy



(b) Effect on over(under)-reliance

Figure B.5 **Effect of explainability ξ on over(under)-reliance and accuracy ($\mu = 0.65, \lambda = 0.5$, $\theta = 0.5$, $\lambda_e = 0.4$).**

In this section, we further extend our model by introducing an additional cognitive cost associated with processing the machine's signals, including its predictions. This cost is assumed to be asymmetric with respect to the explanation signal. Specifically, the DM incurs a cost when updating their belief from μ to μ^{se} only when the explanation indicates high machine quality (i.e., $e = \hbar$). As illustrated in Figure B.2, the DM can more readily conclude that the machine is uninformative when $e = \ell$, without expending much cognitive effort. In contrast, when $e = \hbar$, the DM may need to exert additional effort to interpret and make sense of the explanation provided by the machine.

As in §6.2, we adopt an entropic cost structure and assume that the marginal cost of processing machine signals is lower than the marginal cost of assessing the task itself, denoted by $\lambda_s < \lambda$. The expected cognitive cost from this belief update is then given by:

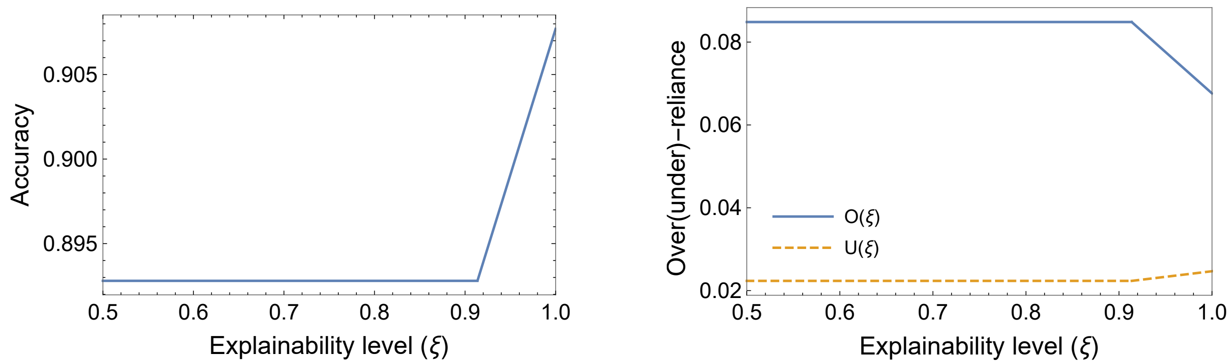
$$C_s(\xi) = \lambda_s \left(H(\mu) - \mathbb{E}\left[H\left(\mu^{S+}\right) | E = \hbar \right] \right).$$

where λ_s increases with the cognitive difficulty of interpreting the machine's signals when $e = \hbar$.

Proposition B.1. *For any $\lambda_s \in (0, \lambda]$, $C_s(\xi)$ is increasing in ξ.*

Proposition B.1 shows that the cognitive cost of interpreting machine signals grows with explainability as higher explainability leads to more informative (and thus more polarizing) updates to the DM's belief, thereby increasing the uncertainty reduction and the associated cost.

This added cognitive burden may lead the DM to forgo the machine entirely. Let $V_s(\xi) = A(\xi) - C(\xi) - C_s(\xi)$ denote the expected value from using the machine and processing its signals, and $V_h^\star(\mu) = A_h^\star(\mu) - C_h^\star(\mu)$ the expected value from relying solely on prior beliefs. The DM then solves: $\max\{V_h^\star(\mu), V_s(\xi)\}$. Since $V_h^\star(\mu)$ is constant in ξ and $V_s(\xi)$ decreases in λ_s (as $C_s(\xi)$ increases), there exists a threshold $\lambda_s^* < \lambda$ such that when $\lambda_s > \lambda_s^*$, the DM strictly prefers to ignore the machine for any $\xi \in [1/2, 1]$.



(a) Effect on accuracy (b) Effect on over(under)-reliance

Figure B.6 Effect of explainability ξ on over(under)-reliance and accuracy when the DM selectively exerts effort ($\mu = 0.25, \lambda = 0.5,\ \theta = 0.5$).

B.5. Illustration of the Reliance Behavior Results in §6.3

Figure B.6 illustrates the impact of explainability when the DM selectively exerts effort when $s = +$.

B.6. Visualization of the Parameter Regions for Cognitive Effort Results

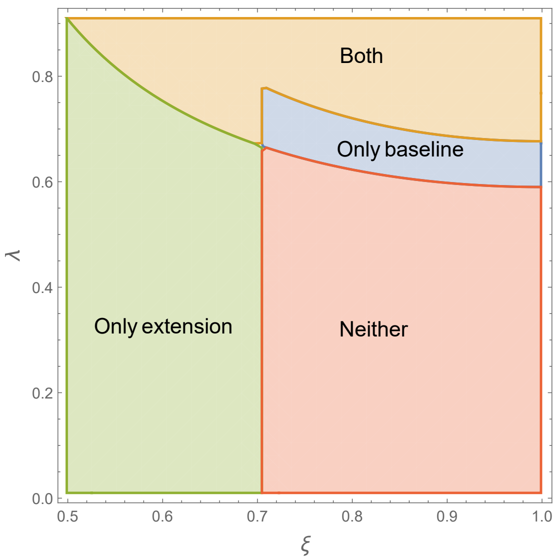
In §6, we extend one of our main results that explainability may increase the DM's cognitive effort along several dimensions. Specifically, Propositions 6, 10, 13, and 15 extend Theorem 3 to settings featuring, respectively, an explainability-accuracy trade-off, costly explanation processing, selective effort by the DM, and asymmetric payoffs. More specifically, all of these results identify sufficient conditions under which there exists a range of explainability levels ξ for which explainability increases cognitive effort relative to the no-explanation benchmark (i.e., $\xi = 1/2$). Figure B.7 illustrates these parameter regions in the two-dimensional $\xi - \lambda$ space and contrasts them with the baseline model. In all figures, the range of the information cost λ shown on the y-axis corresponds to the complementarity conditions specified in Proposition 2 for the baseline case, and to the analogous parameter bounds for the extensions.

Figure B.7a illustrates that, in the presence of a negative explainability-accuracy tradeoff (e.g., $\gamma^{\hbar} = 0.1$ and $\gamma^{\ell} = 0$), cognitive effort is guaranteed to increase over a range of explainability levels regardless of the information cost λ. In contrast, in the baseline model, this increase is guaranteed only when λ is sufficiently large. The latter is also the case when processing explanations requires extra cognitive effort (e.g., $\lambda_e = 0.2$), as Figure B.7b illustrates.

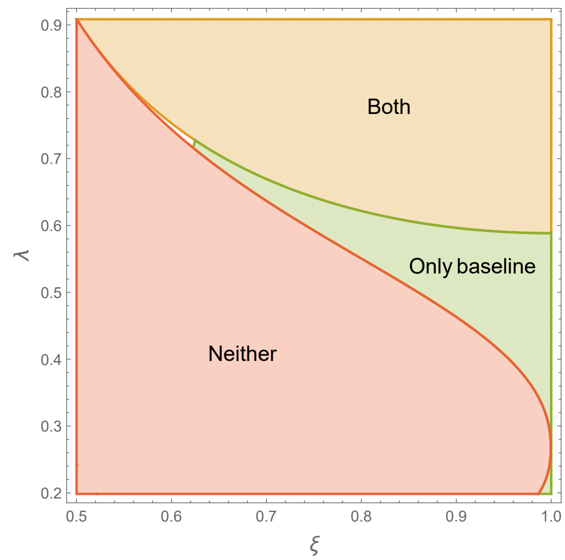
In §6.3, we study a selective effort regime in which the DM processes machine explanations only when the machine provides a positive prediction (i.e., $s = +$). Because we assume $\mu^- \leq \underline{\mu}$, this puts the analysis outside of the complementarity regime we consider in our baseline model, as specified in Proposition 11. Consequently, Figure B.7c is partitioned into two mutually exclusive regions, $[0, \overline{\lambda}]$ and $[\overline{\lambda}, \overline{\lambda}_s]$, corresponding to the parameter regions in which our baseline and extension settings apply, respectively. As also established in Proposition 13, the existence of explainability levels for which cognitive effort is increased is guaranteed when the information cost λ is sufficiently high in both settings.

Finally, Figure B.7d illustrates these regions for our baseline setting as well as our extension scenario where making a false-negative error is costlier than a false-positive one (e.g., $\delta = 1.25$). Consistent with Proposition

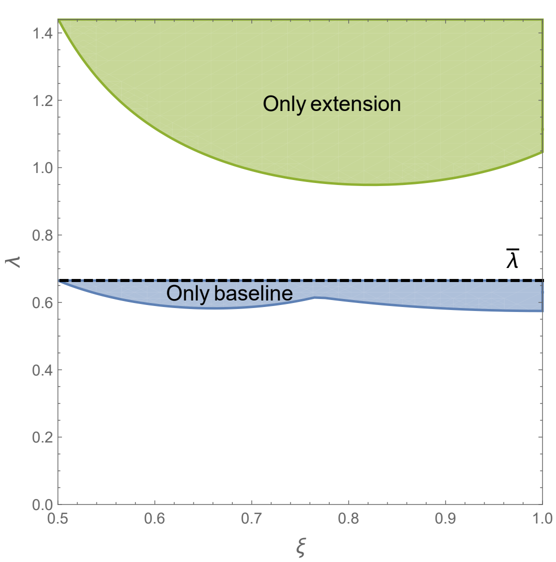
15, when λ is sufficiently large, there exists a parameter region for explainability for which cognitive effort is higher than in the no-explanation benchmark.



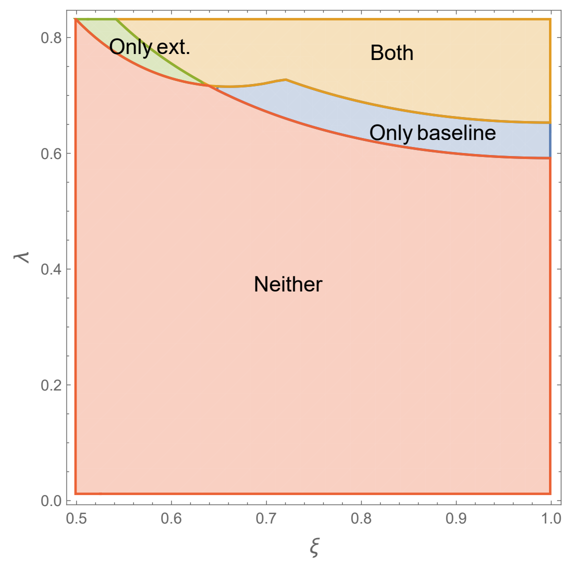
(a) Proposition 6 ($\mu = 0.5, \gamma^{\hbar} = 0.1, \gamma^{\ell} = 0$)



(b) Proposition 10 ($\mu = 0.5, \lambda_e = 0.2$)



(c) Proposition 13 ($\mu = 0.4$)



(d) Proposition 15 ($\mu = 0.5$, $\delta = 1.25$)

Figure B.7 **Parameter regions where cognitive effort is increased in the baseline model and extensions ($\theta = 0.5$)**

References

Matějka F, McKay A (2015) Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review* 105(1):272–98.

Vasconcelos H, Jörke M, Grunde-McLaughlin M, Gerstenberg T, Bernstein MS, Krishna R (2023) Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7(CSCW1):1–38.