Supplementary Material: The AI off-switch problem as a signalling game: bounded rationality and incomparability

Alessio Benavoli[1] **Alessandro Facchini**[2] **Marco Zaffalon**[2]

[1]School of Computer Science and Statistic, Trinity College Dublin, Ireland
[2]SUPSI, IDSIA - Dalle Molle Institute for Artificial Intelligence, Lugano, Switzerland.

A. Useful results

In this section, we have listed some useful results involving Gaussian integrals [2] that we will use in the proofs.

Lemma A.1. *List of useful Gaussian integrals:*

$$\int_{-\infty}^{\infty} \phi(x)\phi(a+bx)\,dx = \frac{1}{\sqrt{1+b^2}}\phi\left(\frac{a}{\sqrt{1+b^2}}\right), \tag{1}$$

$$\int_{-\infty}^{\infty} \Phi(a+bx)\phi(x)\,dx = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right), \tag{2}$$

$$\int_{-\infty}^{\infty} x\Phi(a+bx)\phi(x)\,dx = \frac{b}{\sqrt{1+b^2}}\phi\left(\frac{a}{\sqrt{1+b^2}}\right), \tag{3}$$

where Φ, ϕ are the CDF and, respectively, PDF of a standard normal distribution.

The *inverse Mills ratio* states [1]:

Lemma A.2. *For $x \sim N(m, s^2)$, it states that:*

$$E[xI_{\{a\leq x\leq b\}}] = m\left(\Phi\left(\tfrac{b-m}{s}\right) - \Phi\left(\tfrac{a-m}{s}\right)\right) - s\left(\phi\left(\tfrac{b-m}{s}\right) - \phi\left(\tfrac{a-m}{s}\right)\right). \tag{4}$$

From the above lemma, we can prove:

Lemma A.3. *For $x \sim N(m, s^2)$,*

$$E[|x|] = m\left(1-2\Phi\left(\tfrac{-m}{s}\right)\right) + 2s\phi\left(\tfrac{-m}{s}\right). \tag{5}$$

Proof. Rewrite $|x| = xI_{\{x\geq 0\}} - xI_{\{x<0\}}$ and apply (4):

$$E[xI_{\{x\geq 0\}}] = m\left(1-\Phi\left(\tfrac{-m}{s}\right)\right) - s\left(-\phi\left(\tfrac{-m}{s}\right)\right) \tag{6}$$

and

$$E[xI_{\{x<0\}}] = m\left(\Phi\left(\tfrac{-m}{s}\right)\right) - s\left(\phi\left(\tfrac{-m}{s}\right)\right) \tag{7}$$

and, therefore,

$$E[|x|] = m\left(1-2\Phi\left(\tfrac{-m}{s}\right)\right) + 2s\phi\left(\tfrac{-m}{s}\right). \tag{8}$$

□

Finally, we prove the following main lemma which we will use to prove the results in the paper.

Lemma A.4. *Consider $x, o \in \mathcal{X}$ and assume that*

$$\begin{bmatrix}\nu(x)\\ \nu(o)\end{bmatrix} \sim N\left(\begin{bmatrix}\mu_p(x)\\ \mu_p(o)\end{bmatrix}, \begin{bmatrix}K_p(x,x) & K_p(x,o)\\ K_p(o,x) & K_p(o,o)\end{bmatrix}\right), \tag{9}$$

and $n(x), n(o) \sim N(0, \sigma^2)$ (independent noise). Then we have that

$$\begin{aligned}
&E[\nu(x)I_{\{\nu(x)+n(x)>\nu(o)+n(o)\}}]\\
&= \mu_p(x)\left(1-\Phi\left(\tfrac{(\mu_p(o)-\mu_p(x))}{\sqrt{K_p(x,x)+2\sigma^2+K_p(o,o)-2K_p(x,o)}}\right)\right)\\
&+ \tfrac{K_p(x,x)-K_p(x,o)}{\sqrt{K_p(x,x)+2\sigma^2+K_p(o,o)-2K_p(x,o)}}\\
&\cdot\phi\left(\tfrac{\mu_p(o)-\mu_p(x)}{\sqrt{K_p(x,x)+2\sigma^2+K_p(o,o)-2K_p(x,o)}}\right)
\end{aligned} \tag{10}$$

Proof. We will compute $E[\nu(x)I_{\{\nu(x)>\nu(o)+n(o)-n(x)\}}]$ in two steps. First, we assume that $\nu(o), n(o)-n(x)$ are given and, therefore, we condition the joint PDF of $\nu(x), n(x), \nu(o), n(o)$ on $\nu(o), n(o)$. Since only the variables $\nu(x), \nu(o)$ are dependent, then we have

$$\begin{aligned}
&p(\nu(x)|\nu(o)) =\\
&N\left(\nu(x); \mu_p(x)+\tfrac{K_p(x,o)}{K_p(o,o)}(\nu(o)-\mu_p(o)), K_p(x,x)-\tfrac{K_p^2(x,o)}{K_p(o,o)}\right).
\end{aligned} \tag{11}$$

Therefore, we can apply (4) conditionally on $\nu(o), n(o)-n(x)$ which leads to

$$\begin{aligned}
&E[\nu(x)I_{\{\nu(x)>\nu(o)+n(o)-n(x)\}}|\nu(o), n(o), n(x)]\\
&= m_1\left(1-\Phi\left(\tfrac{\nu(o)+n(o)-n(x)-m_1}{\sigma_1}\right)\right) + \sigma_1\phi\left(\tfrac{\nu(o)+n(o)-n(x)-m_1}{\sigma_1}\right)
\end{aligned} \tag{12}$$

Now observe that

$$\begin{aligned}
E[m_1] = \int &\left(\mu_p(x) + \tfrac{K_p(x,o)}{K_p(o,o)}(\nu(o)-\mu_p(o))\right)\\
&N(\nu(o); \mu_p(o), K_p(o,o))d\nu(o)d\nu(o) = \mu_p(x).
\end{aligned} \tag{13}$$

and

$$
\begin{aligned}
&E\left[m_1\Phi\left(\frac{\nu(o)+n(o)-n(x)-m_1}{\sigma_1}\right)\right]\\
&=E\left[\left(\mu_p(x)+\frac{K_p(x,o)}{K_p(o,o)}(\nu(o)-\mu_p(o))\right)\Phi\left(\frac{\nu(o)+n(o)-n(x)-m_1}{\sigma_1}\right)\right]\\
&=\left(\mu_p(x)-\frac{K_p(x,o)}{K_p(o,o)}\mu_p(o)\right)E\left[\Phi\left(\frac{\nu(o)+n(o)-n(x)-m_1}{\sigma_1}\right)\right]\\
&+\frac{K_p(x,o)}{K_p(o,o)}E\left[\nu(o)\Phi\left(\frac{\nu(o)+n(o)-n(x)-m_1}{\sigma_1}\right)\right]
\end{aligned}
\tag{14}
$$

The expectations are with respect to $\nu(o), n(o), n(x)$. Now we use (2) to get the following result:

$$
\begin{aligned}
&\int\Phi\left(\frac{\nu(o)+n(o)-n(x)-m_1}{\sigma_1}\right)N(n(o)-n(x);0,2\sigma^2)dn(o)\\
&=\Phi\left(\frac{\nu(o)-m_1}{\sqrt{\sigma_1^2+2\sigma^2}}\right),
\end{aligned}
\tag{15}
$$

and so:

$$
\begin{aligned}
&E\left[\Phi\left(\frac{\nu(o)+n(o)-n(x)-m_1}{\sigma_1}\right)\right]\\
&=\int\Phi\left(\frac{\nu(o)-m_1}{\sqrt{\sigma_1^2+2\sigma^2}}\right)N(\nu(o);\mu_p(o),K_p(o,o))d\nu(o)\\
&=\int\Phi\left(\frac{\nu(o)\frac{K_p(o,o)-K_p(x,o)}{K_p(o,o)}+m_2}{\sqrt{\sigma_1^2+2\sigma^2}}\right)\\
&\quad N(\nu(o);\mu_p(o),K_p(o,o))d\nu(o)\\
&=\int\Phi\left(\frac{z\frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(o,o)}}+m_2+\frac{K_p(o,o)-K_p(x,o)}{K_p(o,o)}\mu_p(o)}{\sqrt{\sigma_1^2+2\sigma^2}}\right)\\
&\quad N(z;0,1)dz\\
&=\Phi\left(\frac{\mu_p(o)\frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(o,o)}}+m_2\sqrt{K_p(o,o)}}{\sqrt{K_p(o,o)(\sigma_1^2+2\sigma^2)+(K_p(o,o)-K_p(x,o))^2}}\right)\\
&=\Phi\left(\frac{\sqrt{K_p(o,o)}(\mu_p(o)-\mu_p(x))}{\sqrt{K_p(o,o)(\sigma_1^2+2\sigma^2)+(K_p(o,o)-K_p(x,o))^2}}\right),
\end{aligned}
\tag{16}
$$

with $m_2=\frac{-K_p(o,o)\mu_p(x)+K_p(x,o)\mu_p(o)}{K_p(o,o)}$. Similarly, we have that

$$
\begin{aligned}
&E\left[\nu(o)\Phi\left(\frac{\nu(o)+n(o)-n(x)-m_1}{\sigma_1}\right)\right]\\
&=\int\nu(o)\Phi\left(\frac{\nu(o)-m_1}{\sqrt{\sigma_1^2+2\sigma^2}}\right)N(\nu(o);\mu_p(o),K_p(o,o))d\nu(o)\\
&=\int\nu(o)\Phi\left(\frac{\nu(o)\frac{K_p(o,o)-K_p(x,o)}{K_p(o,o)}+m_2}{\sqrt{\sigma_1^2+2\sigma^2}}\right)\\
&\quad N(\nu(o);\mu_p(o),K_p(o,o))d\nu(o)\\
&=\int\Phi\left(\frac{z\frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(o,o)}}+m_2+\frac{K_p(o,o)-K_p(x,o)}{K_p(o,o)}\mu_p(o)}{\sqrt{\sigma_1^2+2\sigma^2}}\right)\\
&\quad\left(z\sqrt{K_p(o,o)}+\mu_p(o)\right)N(z;0,1)dz
\end{aligned}
\tag{17}
$$

We separate the sum:

$$
\begin{aligned}
&\int\Phi\left(\frac{z\frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(o,o)}}+m_2+\frac{K_p(o,o)-K_p(x,o)}{K_p(o,o)}\mu_p(o)}{\sqrt{\sigma_1^2+2\sigma^2}}\right)\\
&\quad\mu_p(o)N(z;0,1)dz\\
&=\mu_p(o)\Phi\left(\frac{\mu_p(o)\frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(o,o)}}+m_2\sqrt{K_p(o,o)}}{\sqrt{K_p(o,o)(\sigma_1^2+2\sigma^2)+(K_p(o,o)-K_p(x,o))^2}}\right)\\
&=\mu_p(o)\Phi\left(\frac{\sqrt{K_p(o,o)}(\mu_p(o)-\mu_p(x))}{\sqrt{K_p(o,o)(\sigma_1^2+2\sigma^2)+(K_p(o,o)-K_p(x,o))^2}}\right).
\end{aligned}
\tag{18}
$$

The other term in the sum

$$
\begin{aligned}
&\int\Phi\left(\frac{z\frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(o,o)}}+m_2+\frac{K_p(o,o)-K_p(x,o)}{K_p(o,o)}\mu_p(o)}{\sqrt{\sigma_1^2+2\sigma^2}}\right)\\
&\quad z\sqrt{K_p(o,o)}N(z;0,1)dz\\
&=\frac{\sqrt{K_p(o,o)}(K_p(o,o)-K_p(x,o))}{\sqrt{K_p(o,o)(\sigma_1^2+\sigma^2)+(K_p(o,o)-K_p(x,o))^2}}\\
&\quad\phi\left(\frac{\mu_p(o)\frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(o,o)}}+m_2\sqrt{K_p(o,o)}}{\sqrt{K_p(o,o)(\sigma_1^2+2\sigma^2)+(K_p(o,o)-K_p(x,o))^2}}\right)\\
&=\frac{\sqrt{K_p(o,o)}(K_p(o,o)-K_p(x,o))}{\sqrt{K_p(o,o)(\sigma_1^2+2\sigma^2)+(K_p(o,o)-K_p(x,o))^2}}\\
&\quad\phi\left(\frac{\sqrt{K_p(o,o)}(\mu_p(o)-\mu_p(x))}{\sqrt{K_p(o,o)(\sigma_1^2+2\sigma^2)+(K_p(o,o)-K_p(x,o))^2}}\right).
\end{aligned}
\tag{19}
$$

where we have used (3). Finally, we consider

$$\int \phi\left(\frac{\nu(o)+n(o)-n(x)-m_1}{\sigma_1}\right) N(n(o)-n(x);0,2\sigma^2)dn(o) = \frac{\sigma_1}{\sqrt{\sigma_1^2+2\sigma^2}}\phi\left(\frac{\nu(o)-m_1}{\sqrt{\sigma_1^2+2\sigma^2}}\right), \tag{20}$$

where the last equality follows by (1). We use (16) to get:

$$\int \frac{\sigma_1}{\sqrt{\sigma_1^2+2\sigma^2}}\phi\left(\frac{\nu(o)-m_1}{\sqrt{\sigma_1^2+2\sigma^2}}\right) N(\nu(o);\mu_p(o),K_p(o,o))d\nu(o)$$
$$= \int \frac{\sigma_1}{\sqrt{\sigma_1^2+2\sigma^2}}\phi\left(\frac{\nu(o)\frac{K_p(o,o)-K_p(x,o)}{K_p(o,o)}+m_2}{\sqrt{\sigma_1^2+2\sigma^2}}\right) N(\nu(o);\mu_p(o),K_p(o,o))d\nu(o)$$
$$= \int \frac{\sigma_1}{\sqrt{\sigma_1^2+2\sigma^2}}\phi\left(\frac{z\frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(o,o)}}+m_2+\frac{K_p(o,o)-K_p(x,o)}{K_p(o,o)}\mu_p(o)}{\sqrt{\sigma_1^2+2\sigma^2}}\right) N(z;0,1)d\nu(o)$$
$$= \frac{\sqrt{K_p(o,o)}\sigma_1}{\sqrt{K_p(o,o)(\sigma_1^2+2\sigma^2)+(K_p(o,o)-K_p(x,o))^2}} \phi\left(\frac{\sqrt{K_p(o,o)}(\mu_p(o)-\mu_p(x))}{\sqrt{K_p(o,o)(\sigma_1^2+2\sigma^2)+(K_p(o,o)-K_p(x,o))^2}}\right). \tag{21}$$

Therefore, from (12) and (16)–(21), we obtain

$$E[(\nu(x)+n(x))I_{\{\nu(x)+n(x)>\nu(o)+n(o)\}}] = \mu_p(x)$$
$$-\left(\mu_p(x)-\frac{K_p(x,o)}{K_p(o,o)}\mu_p(o)\right) \cdot \Phi\left(\frac{\sqrt{K_p(o,o)}(\mu_p(o)-\mu_p(x))}{\sqrt{K_p(o,o)(\sigma_1^2+2\sigma^2)+(K_p(o,o)-K_p(x,o))^2}}\right)$$
$$-\frac{K_p(x,o)}{K_p(o,o)}\mu_p(o)\Phi\left(\frac{\sqrt{K_p(o,o)}(\mu_p(o)-\mu_p(x))}{\sqrt{K_p(o,o)(\sigma_1^2+2\sigma^2)+(K_p(o,o)-K_p(x,o))^2}}\right)$$
$$-\frac{K_p(x,o)}{K_p(o,o)}\frac{\sqrt{K_p(o,o)}(K_p(o,o)-K_p(x,o))}{\sqrt{K_p(o,o)(\sigma_1^2+2\sigma^2)+(K_p(o,o)-K_p(x,o))^2}} \phi\left(\frac{\sqrt{K_p(o,o)}(\mu_p(o)-\mu_p(x))}{\sqrt{K_p(o,o)(\sigma_1^2+2\sigma^2)+(K_p(o,o)-K_p(x,o))^2}}\right)$$
$$+\frac{\sqrt{K_p(o,o)}\sigma_1^2}{\sqrt{K_p(o,o)(\sigma_1^2+2\sigma^2)+(K_p(o,o)-K_p(x,o))^2}} \phi\left(\frac{\sqrt{K_p(o,o)}(\mu_p(o)-\mu_p(x))}{\sqrt{K_p(o,o)(\sigma_1^2+2\sigma^2)+(K_p(o,o)-K_p(x,o))^2}}\right)$$
$$= \mu_p(x)\left(1-\Phi\left(\frac{\sqrt{K_p(o,o)}(\mu_p(o)-\mu_p(x))}{\sqrt{K_p(o,o)(\sigma_1^2+2\sigma^2)+(K_p(o,o)-K_p(x,o))^2}}\right)\right)$$
$$+\frac{\sqrt{K_p(o,o)}(K_p(x,x)-K_p(x,o))}{\sqrt{K_p(o,o)(\sigma_1^2+2\sigma^2)+(K_p(o,o)-K_p(x,o))^2}} \phi\left(\frac{\sqrt{K_p(o,o)}(\mu_p(o)-\mu_p(x))}{\sqrt{K_p(o,o)(\sigma_1^2+2\sigma^2)+(K_p(o,o)-K_p(x,o))^2}}\right) \tag{22}$$

Note that

$$K_p(o,o)(\sigma_1^2+2\sigma^2)+(K_p(o,o)-K_p(x,o))^2$$
$$= K_p(o,o)\left(K_p(x,x)-\frac{K_p^2(x,o)}{K_p(o,o)}+2\sigma^2\right) + (K_p(o,o)-K_p(x,o))^2$$
$$= K_p(o,o)K_p(x,x)-K_p^2(x,o)+2\sigma^2K_p(o,o) + K_p^2(o,o)+K_p^2(x,o)-2K_p(x,o)K_p(o,o)$$
$$= K_p(o,o)(K_p(x,x)+2\sigma^2+K_p(o,o)-2K_p(x,o)). \tag{23}$$

Therefore, we have that

$$E[\nu(x)I_{\{\nu(x)+n(x)>\nu(o)+n(o)\}}] = \; = \mu_p(x)\left(1-\Phi\left(\frac{(\mu_p(o)-\mu_p(x))}{\sqrt{K_p(x,x)+2\sigma^2+K_p(o,o)-2K_p(x,o)}}\right)\right) + \frac{K_p(x,x)-K_p(x,o)}{\sqrt{K_p(x,x)+2\sigma^2+K_p(o,o)-2K_p(x,o)}} \phi\left(\frac{\mu_p(o)-\mu_p(x)}{\sqrt{K_p(x,x)+2\sigma^2+K_p(o,o)-2K_p(x,o)}}\right) \tag{24}$$

□

B. Proofs

We now move on to the main results.

Proof of Lemma 4.1. The expected value for DEF follows from Lemma A.4 by summing $E[\nu(x)I_{\{\nu(x)+n(x)>\nu(o)+n(o)\}}]$ and $E[\nu(o)I_{\{\nu(x)+n(x)<\nu(o)+n(o)\}}]$. The expected values for IMM, DoN are straightforward.

Proof of Proposition 4.1. The results follow from Lemma 4.1 by considering whether or not the limits $K_0(x,x), K_0(o,o), K_0(o,x) \to 0$ and $\sigma \to 0$ are taken. We make the assumption that whenever $K_o(x,x), K_o(o,o), K_o(o,x) \to 0$ it implies that $K_p(x,x), K_p(o,o), K_p(o,x) \to 0$ (a-priori we have a Dirac's delta). Since K_p depends on both K_0 and σ, we always take the limit with respect to σ first.

If S is **rational** and R has **no uncertainty**, then the expected payoffs can be computed from (17), (18) and (19). The values are $E[DEF] = \max(\mu_p(x), \mu_p(o))$ $E[IMM] = \mu_p(x)$ and $E[DoN] = \mu_p(o)$. Therefore, DEF is never dominated.

If S is **bounded-rational** and R has **no uncertainty**, then $\sigma > 0$ and the payoffs are: $E[DEF] = p\mu_p(x) + (1-p)\mu_p(o)$, $E[IMM] = \mu_p(x)$ and $E[DoN] = \mu_p(o)$, where $p = \Phi\left(\frac{\mu_p(x)-\mu_p(o)}{\sqrt{2\sigma^2}}\right)$ Therefore, $p \in (0,1)$ and DEF is never optimal.

If S is **rational** and R has **uncertainty**, then $E[DEF] = p\mu_p(x) + (1-p)\mu_p(o) + e$, $E[IMM] = \mu_p(x)$ and $E[DoN] = \mu_p(o)$, where $p = \Phi\left(\frac{\mu_p(x)-\mu_p(o)}{\sqrt{K_p(o,o)+K_p(x,x)-2K_p(x,o)}}\right)$ and

$$e = \frac{K_p(x,x)-K_p(x,o)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\phi\left(\frac{\mu_p(o)-\mu_p(x)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right) + \frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\phi\left(\frac{\mu_p(x)-\mu_p(o)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right).$$

When S is rational, then

$$u_R(t, DEF, b^*(DEF)) = \nu(o)I_{\{\nu(o)>\nu(x)\}} + \nu(x)I_{\{\nu(x)>\nu(o)\}} = \max(\nu(o), \nu(x)) \tag{25}$$

max is a convex function and, therefore, by Jensen's inequality $E[\max(\nu(o), \nu(x))] \geq \max(E[\nu(o)], E[\nu(x)])$. Therefore, $p\mu_p(x) + (1-p)\mu_p(o) + e \geq \max(\mu_p(x), \mu_p(o))$ and DEF is always optimal.

The last case follows directly from Lemma 4.1.

Proof of Corollary 4.1. The results follow from Proposition 4.1 (and (5)) after subtracting β to the expected payoff for DEF, where

$$\beta = \gamma E[|\nu(o)|] = \gamma\mu_p(o)\left(1 - 2\Phi\left(\frac{-\mu_p(o)}{\sqrt{K_p(o,o)}}\right)\right) + 2\gamma\sqrt{K_p(o,o)}\phi\left(\frac{-\mu_p(o)}{\sqrt{K_p(o,o)}}\right). \tag{26}$$

If S is **rational** and R has **no uncertainty**, then the expected payoffs can be computed from (17), (18) and (19). The values are $E[DEF] = \max(\mu_p(x), \mu_p(o)) - \gamma'|\mu_p(o)|$ $E[IMM] = \mu_p(x)$ and $E[DoN] = \mu_p(o)$. Therefore, DEF is never optimal.

If S is **bounded-rational** and R has **no uncertainty**, then $\sigma > 0$ and the payoffs are: $E[DEF] = p\mu_p(x) + (1-p)\mu_p(o) - \gamma'|\mu_p(o)|$ $E[IMM] = \mu_p(x)$ and $E[DoN] = \mu_p(o)$, where $p = \Phi\left(\frac{\mu_p(x)-\mu_p(o)}{\sqrt{2\sigma^2}}\right)$ Therefore, $p \in (0,1)$ and DEF is never optimal.

If S is **rational** and R has **uncertainty**, then $E[DEF] = p\mu_p(x) + (1-p)\mu_p(o) + e - \gamma'|\mu_p(o)|$ $E[IMM] = \mu_p(x)$ and $E[DoN] = \mu_p(o)$. Therefore, DEF is optimal if $p\mu_p(x) + (1-p)\mu_p(o) + e - \gamma'|\mu_p(o)| \geq \max(\mu_p(x), \mu_p(o))$. The last case follows similarly from Proposition 4.1

Proof of Lemma 4.2. The expected value for DEF is equal to the sum of $E[\nu(x)I_{\{\nu(x)>\nu(o)+\sigma)\}}]$, $E[\nu(o)I_{\{\nu(o)>\nu(x)+\sigma)\}}]$ and $\{E[\nu(x)I_{\{|\nu(x)-\nu(o)|\leq\sigma)\}}], E[\nu(o)I_{\{|\nu(x)-\nu(o)|\leq\sigma)\}}]\}$. For fixed $\nu(x)$, we have that

$$p(\nu(x)|\nu(o)) = N(\nu(x); m_1, \sigma_1^2) = N\left(\nu(x); \mu_p(x) + \frac{K_p(x,o)}{K_p(o,o)}(\nu(o)-\mu_p(o)), K_p(x,x) - \frac{K_p^2(x,o)}{K_p(o,o)}\right). \tag{27}$$

Therefore, we can apply (4) conditionally on $\nu(o)$ which leads to

$$E[\nu(x)I_{\{\nu(x)>\nu(o)+\sigma\}}|\nu(o)] = m_1\left(1 - \Phi\left(\frac{\nu(o)+\sigma-m_1}{\sigma_1}\right)\right) + \sigma_1\phi\left(\frac{\nu(o)+\sigma-m_1}{\sigma_1}\right) \tag{28}$$

Now observe that

$$E[m_1] = \int \left(\mu_p(x) + \frac{K_p(x,o)}{K_p(o,o)}(\nu(o) - \mu_p(o))\right) N(\nu(o); \mu_p(o), K_p(o,o))d\nu(o)d\nu(o) = \mu_p(x), \tag{29}$$

and

$$\begin{aligned} &E\left[m_1\Phi\left(\frac{\nu(o)+\sigma-m_1}{\sigma_1}\right)\right] \\ &= E\left[\left(\mu_p(x) + \frac{K_p(x,o)}{K_p(o,o)}(\nu(o) - \mu_p(o))\right)\Phi\left(\frac{\nu(o)+\sigma-m_1}{\sigma_1}\right)\right] \\ &= \left(\mu_p(x) - \frac{K_p(x,o)}{K_p(o,o)}\mu_p(o)\right)E\left[\Phi\left(\frac{\nu(o)+\sigma-m_1}{\sigma_1}\right)\right] \\ &+ \frac{K_p(x,o)}{K_p(o,o)}E\left[\nu(o)\Phi\left(\frac{\nu(o)+\sigma-m_1}{\sigma_1}\right)\right] \end{aligned} \tag{30}$$

The expectations are with respect to $\nu(o)$. Now we use (2) to get

$$
\begin{aligned}
&E\left[\Phi\left(\frac{\nu(o)+\sigma-m_1}{\sigma_1}\right)\right]\\
&=\int \Phi\left(\frac{\nu(o)+\sigma-m_1}{\sigma_1}\right)N(\nu(o);\mu_p(o),K_p(o,o))d\nu(o)\\
&=\int \Phi\left(\frac{\nu(o)\frac{K_p(o,o)-K_p(x,o)}{K_p(o,o)}+m_2}{\sigma_1}\right)\\
&\qquad N(\nu(o);\mu_p(o),K_p(o,o))d\nu(o)\\
&=\int \Phi\left(\frac{z\frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(o,o)}}+m_2+\frac{K_p(o,o)-K_p(x,o)}{K_p(o,o)}\mu_p(o)}{\sigma_1}\right)\\
&\qquad N(z;0,1)dz\\
&=\Phi\left(\frac{\mu_p(o)\frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(o,o)}}+m_2\sqrt{K_p(o,o)}}{\sqrt{K_p(o,o)\sigma_1^2+(K_p(o,o)-K_p(x,o))^2}}\right)\\
&=\Phi\left(\frac{\sqrt{K_p(o,o)}(\mu_p(o)+\sigma-\mu_p(x))}{\sqrt{K_p(o,o)\sigma_1^2+(K_p(o,o)-K_p(x,o))^2}}\right),
\end{aligned}
\tag{31}
$$

with $m_2=\frac{K_p(o,o)(\sigma-\mu_p(x))+K_p(x,o)\mu_p(o)}{K_p(o,o)}$. Similarly, we have that

$$
\begin{aligned}
&E\left[\nu(o)\Phi\left(\frac{\nu(o)+\sigma-m_1}{\sigma_1}\right)\right]\\
&=\int \nu(o)\Phi\left(\frac{\nu(o)+\sigma-m_1}{\sigma_1}\right)N(\nu(o);\mu_p(o),K_p(o,o))d\nu(o)\\
&=\int \nu(o)\Phi\left(\frac{\nu(o)\frac{K_p(o,o)-K_p(x,o)}{K_p(o,o)}+m_2}{\sigma_1}\right)\\
&\qquad N(\nu(o);\mu_p(o),K_p(o,o))d\nu(o)\\
&=\int \Phi\left(\frac{z\frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(o,o)}}+m_2+\frac{K_p(o,o)-K_p(x,o)}{K_p(o,o)}\mu_p(o)}{\sigma_1}\right)\\
&\left(z\sqrt{K_p(o,o)}+\mu_p(o)\right)N(z;0,1)dz
\end{aligned}
\tag{32}
$$

We separate the sum:

$$
\begin{aligned}
&\int \Phi\left(\frac{z\frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(o,o)}}+m_2+\frac{K_p(o,o)-K_p(x,o)}{K_p(o,o)}\mu_p(o)}{\sigma_1}\right)\\
&\mu_p(o)N(z;0,1)dz\\
&=\mu_p(o)\Phi\left(\frac{\mu_p(o)\frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(o,o)}}+m_2\sqrt{K_p(o,o)}}{\sqrt{K_p(o,o)\sigma_1^2+(K_p(o,o)-K_p(x,o))^2}}\right)\\
&=\mu_p(o)\Phi\left(\frac{\sqrt{K_p(o,o)}(\mu_p(o)+\sigma-\mu_p(x))}{\sqrt{K_p(o,o)\sigma_1^2+(K_p(o,o)-K_p(x,o))^2}}\right).
\end{aligned}
\tag{33}
$$

The other term in the sum

$$
\begin{aligned}
&\int \Phi\left(\frac{z\frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(o,o)}}+m_2+\frac{K_p(o,o)-K_p(x,o)}{K_p(o,o)}\mu_p(o)}{\sigma_1}\right)\\
&z\sqrt{K_p(o,o)}N(z;0,1)dz\\
&=\frac{\sqrt{K_p(o,o)}(K_p(o,o)-K_p(x,o))}{\sqrt{K_p(o,o)\sigma_1^2+(K_p(o,o)-K_p(x,o))^2}}\\
&\phi\left(\frac{\mu_p(o)\frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(o,o)}}+m_2\sqrt{K_p(o,o)}}{\sqrt{K_p(o,o)(\sigma_1^2+2\sigma^2)+(K_p(o,o)-K_p(x,o))^2}}\right)\\
&=\frac{\sqrt{K_p(o,o)}(K_p(o,o)-K_p(x,o))}{\sqrt{K_p(o,o)\sigma_1^2+(K_p(o,o)-K_p(x,o))^2}}\\
&\phi\left(\frac{\sqrt{K_p(o,o)}(\mu_p(o)+\sigma-\mu_p(x))}{\sqrt{K_p(o,o)\sigma_1^2+(K_p(o,o)-K_p(x,o))^2}}\right).
\end{aligned}
\tag{34}
$$

where we have used (3). Finally, we consider

$$
\begin{aligned}
&\int \phi\left(\frac{\nu(o)+\sigma-m_1}{\sigma_1}\right)N(\nu(o);\mu_p(o),K_p(o,o))d\nu(o)\\
&=\int \phi\left(\frac{z\frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(o,o)}}+m_2+\frac{K_p(o,o)-K_p(x,o)}{K_p(o,o)}\mu_p(o)}{\sigma_1}\right)\\
&N(z;0,1)d\nu(o)\\
&=\frac{\sqrt{K_p(o,o)}\sigma_1}{\sqrt{K_p(o,o)_p\sigma_1^2+(K_p(o,o)-K_p(x,o))^2}}\\
&\phi\left(\frac{\sqrt{K_p(o,o)}(\mu_p(o)+\sigma-\mu_p(x))}{\sqrt{K_p(o,o)\sigma_1^2+(K_p(o,o)-K_p(x,o))^2}}\right).
\end{aligned}
\tag{35}
$$

Therefore, from (12) and (16)–(21), we obtain

$$
\begin{aligned}
&E[\nu(x)I_{\{\nu(x)>\nu(o)+\sigma\}}] = \mu_p(x)\\
&-\left(\mu_p(x) - \tfrac{K_p(x,o)}{K_p(o,o)}\mu_p(o)\right)\\
&\cdot \Phi\left(\frac{\sqrt{K_p(o,o)}(\mu_p(o)+\sigma-\mu_p(x))}{\sqrt{K_p(o,o)\sigma_1^2+(K_p(o,o)-K_p(x,o))^2}}\right)\\
&-\tfrac{K_p(x,o)}{K_p(o,o)}\mu_p(o)\Phi\left(\frac{\sqrt{K_p(o,o)}(\mu_p(o)+\sigma-\mu_p(x))}{\sqrt{K_p(o,o)\sigma_1^2+(K_p(o,o)-K_p(x,o))^2}}\right)\\
&-\tfrac{K_p(x,o)}{K_p(o,o)}\frac{\sqrt{K_p(o,o)}(K_p(o,o)-K_p(x,o))}{\sqrt{K_p(o,o)\sigma_1^2+(K_p(o,o)-K_p(x,o))^2}}\\
&\phi\left(\frac{\sqrt{K_p(o,o)}(\mu_p(o)+\sigma-\mu_p(x))}{\sqrt{K_p(o,o)\sigma_1^2+(K_p(o,o)-K_p(x,o))^2}}\right)\\
&+\frac{\sqrt{K_p(o,o)}\sigma_1^2}{\sqrt{K_p(o,o)\sigma_1^2+(K_p(o,o)-K_p(x,o))^2}}\\
&\phi\left(\frac{\sqrt{K_p(o,o)}(\mu_p(o)+\sigma-\mu_p(x))}{\sqrt{K_p(o,o)\sigma_1^2+(K_p(o,o)-K_p(x,o))^2}}\right)\\
&=\mu_p(x)\left(1-\Phi\left(\frac{\sqrt{K_p(o,o)}(\mu_p(o)+\sigma-\mu_p(x))}{\sqrt{K_p(o,o)\sigma_1^2+(K_p(o,o)-K_p(x,o))^2}}\right)\right)\\
&+\frac{\sqrt{K_p(o,o)}(K_p(x,x)-K_p(x,o))}{\sqrt{K_p(o,o)(\sigma_1^2+2\sigma^2)+(K_p(o,o)-K_p(x,o))^2}}\\
&\phi\left(\frac{\sqrt{K_p(o,o)}(\mu_p(o)+\sigma-\mu_p(x))}{\sqrt{K_p(o,o)\sigma_1^2+(K_p(o,o)-K_p(x,o))^2}}\right)
\end{aligned}
\tag{36}
$$

Note that

$$
\begin{aligned}
&K_p(o,o)\sigma_1^2 + (K_p(o,o)-K_p(x,o))^2\\
&=K_p(o,o)\left(K_p(x,x)-\tfrac{K_p^2(x,o)}{K_p(o,o)}\right)+(K_p(o,o)-K_p(x,o))^2\\
&=K_p(o,o)K_p(x,x)-K_p^2(x,o)\\
&+K_p^2(o,o)+K_p^2(x,o)-2K_p(x,o)K_p(o,o)\\
&=K_p(o,o)(K_p(x,x)+K_p(o,o)-2K_p(x,o)).
\end{aligned}
\tag{37}
$$

Therefore, we have that

$$
\begin{aligned}
&E[\nu(x)I_{\{\nu(x)>\nu(o)+\sigma\}}] =\\
&=\mu_p(x)\left(1-\Phi\left(\frac{(\mu_p(o)+\sigma-\mu_p(x))}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right)\right)\\
&+\frac{K_p(x,x)-K_p(x,o)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\\
&\phi\left(\frac{\mu_p(o)+\sigma-\mu_p(x)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right)
\end{aligned}
\tag{38}
$$

and

$$
\begin{aligned}
&E[\nu(o)I_{\{\nu(o)>\nu(x)+\sigma\}}] =\\
&=\mu_p(o)\left(1-\Phi\left(\frac{(\mu_p(x)+\sigma-\mu_p(o))}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right)\right)\\
&+\frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\\
&\phi\left(\frac{\mu_p(x)+\sigma-\mu_p(o)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right)
\end{aligned}
\tag{39}
$$

The other two terms are:

$$
\begin{aligned}
&E[\nu(o)I_{\{|\nu(o)-\nu(x)|\le\sigma\}}] = \mu_p(o)\\
&-E[\nu(o)I_{\{\nu(o)>\nu(x)+\sigma\}}]-E[\nu(o)I_{\{\nu(x)>\nu(o)+\sigma\}}]\\
&=\mu_p(o)-\mu_p(o)\left(1-\Phi\left(\frac{(\mu_p(x)+\sigma-\mu_p(o))}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right)\right)\\
&-\frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\\
&\phi\left(\frac{\mu_p(x)+\sigma-\mu_p(o)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right)\\
&-\mu_p(o)\left(1-\Phi\left(\frac{(-\mu_p(x)+\sigma+\mu_p(o))}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right)\right)\\
&+\frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\\
&\phi\left(\frac{-\mu_p(x)+\sigma+\mu_p(o)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right)\\
&=\mu_p(o)-\mu_p(o)\Bigg(2-\Phi\left(\frac{(\mu_p(x)+\sigma-\mu_p(o))}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right)\\
&-\Phi\left(\frac{(-\mu_p(x)+\sigma+\mu_p(o))}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right)\Bigg)\\
&+\frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\\
&\Bigg(\phi\left(\frac{-\mu_p(x)+\sigma+\mu_p(o)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right)\\
&-\phi\left(\frac{\mu_p(x)+\sigma-\mu_p(o)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right)\Bigg)
\end{aligned}
\tag{40}
$$

and

$$E[\nu(x)I_{\{|\nu(o)-\nu(x)|\leq\sigma\}}] = \\ = \mu_p(x) - \mu_p(x)\Bigg(2 - \Phi\left(\frac{(\mu_p(x)+\sigma-\mu_p(o))}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right) \\ - \Phi\left(\frac{(-\mu_p(x)+\sigma+\mu_p(o))}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right)\Bigg) \\ + \frac{K_p(x,o)-K_p(x,o)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}} \\ \left(\phi\left(\frac{-\mu_p(o)+\sigma+\mu_p(x)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right) \\ - \phi\left(\frac{\mu_p(o)+\sigma-\mu_p(x)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right)\right) \tag{41}$$

Proof of Proposition 4.2. If R has **no uncertainty** and S is rational, then the expected payoffs can be computed from (23), (24) and (25). The values are $E[DEF] = \max(\mu_p(x), \mu_p(o))$, $E[IMM] = \mu_p(x)$ and $E[DoN] = \mu_p(o)$. Therefore, DEF is always optimal.

If S is **bounded-rational** and R has **no uncertainty**, then $\sigma > 0$. We consider three cases: (1) $\mu_p(x) > \mu_p(o)+\sigma$; (2) $\mu_p(o) > \mu_p(x) + \sigma$; (3) otherwise.

In case (1), the payoffs are: $E[DEF] = \mu_p(x)$, $E[IMM] = \mu_p(x)$ and $E[DoN] = \mu_p(o)$, Therefore, DEF is optimal. A similar results holds in case (2). In case (3), $E[DEF] = \{\mu_p(x) - \epsilon, \mu_p(o) - \epsilon\}$. Under the condition (A) or (B), DEF will alwys be dominated. If S is **rational** and R has **uncertainty**, then DEF is optimal as in Proposition 4.1 The last case follows directly from Lemma 4.2.

Proof of Proposition 4.3. The only case where the content of the message is important is when DEF is not optimal. In this case, R makes a decision autonomously.

Whenever DEF is not optimal, the best action can be either IMM(x) if $\mu_p(x) > \mu_p(o)$ or DoN if $\mu_p(o) > \mu_p(x)$. Therefore, if S sends a biased message such that R estimates $\mu_p(x) > \mu_p(o)$ when, in reality, $\nu(x) < \nu(o)$, then R would choose an action that is not optimal for S.

Proof of Proposition 5.1. This follows from Proposition 4.1 for the cases when $\boldsymbol{\nu}(o)$ and $\boldsymbol{\nu}(x)$ are comparable (i.e., one dominates the other). If S is **rational** and R has **no uncertainty**, then if $\boldsymbol{\nu}(o)$ and $\boldsymbol{\nu}(x)$ are comparable, the payoff for DEF(x) will be the best between $\boldsymbol{\nu}(o)$ and $\boldsymbol{\nu}(x)$ and, therefore DEF is not dominated.

If S is **bounded-rational** and R has **no uncertainty**, then if $\boldsymbol{\nu}(o)$ and $\boldsymbol{\nu}(x)$ are comparable, the payoff for DEF(x) will never be optimal, because it will be $p\boldsymbol{\nu}(o) + (1-p)\boldsymbol{\nu}(x)$. If S is **rational** and R has **uncertainty**, then if $\boldsymbol{\nu}(o)$ and $\boldsymbol{\nu}(x)$ are comparable, the payoff for DEF(x) will always be optimal as shown in Proposition 4.1. If S is **bounded-rational** and R has **uncertainty**, the best decision depends on the specific case.

References

[1] G. Grimmett and S. Stirzaker. *Probability Theory and Random Processes*. 3rd. Cambridge University Press, 2001.

[2] Donald Bruce Owen. "A table of normal integrals: A table". In: *Communications in Statistics-Simulation and Computation* 9.4 (1980), pp. 389–419.