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

Abstract

The off-switch problem is a critical challenge in AI control: if an AI system resists being switched off, it poses a significant risk. In this paper, we model the off-switch problem as a signalling game, where a human decision-maker communicates its preferences about some underlying decision problem to an AI agent, which then selects actions to maximise the human's utility. We assume that the human is a bounded rational agent and explore various bounded rationality mechanisms. Using real machine learning models, we reprove prior results and demonstrate that a necessary condition for an AI system to refrain from disabling its off-switch is its uncertainty about the human's utility. We also analyse how message costs influence optimal strategies and extend the analysis to scenarios involving incomparability.

Keywords. off-switch, AI alignment, utility, incomparability

1. Introduction

With the rapid advancements in AI, the challenge of designing systems that are beneficial to humans is becoming increasingly critical. In the rest of the paper, we will refer to an AI system as a *robot* and assume that it is an agent seeking to maximise a utility function ν. A central concern in AI safety is ensuring that the utility function ν aligns with human interests. If a robot's goals conflict with human values, it could make harmful or even adversarial decisions. Moreover, the robot might resist any attempts by its human creators to modify its utility function or allow them to switch it off

To ensure alignment with human values, from the perspective of utilitarianism[1], robots should be designed to maximise humans' utilities (preferences) rather than pursue their own goals. This led [24] to introduce the concept of *AI assistance game*, where an AI agent simply becomes a personal AI assistant. In this setting, AI alignment can be formalised through three key principles [24]:

P1: The Robot's only objective is to maximise the realisation of human preferences.

P2: The Robot is initially uncertain about what those preferences are.

P3: The ultimate source of information about human preferences is human behaviour.

However, designing a Robot to follow these principles does not guarantee that we will be able to control it. The ability to control the Robot ultimately depends on our capacity to switch it off. Will the Robot allow that? The off-switch problem lies at the heart of the control challenge for AI systems. If a machine cannot be switched off because it actively prevents us from doing so, it poses a serious threat. In a seminal paper [17] formulated the off-switch problem as a game between a Robot (R) and a Human (S for Sapiens). R has three possible actions: (1) take an immediate decision (about some underlying decision problem); (2) defer the decision to S; (3) do nothing. If R selects actions (2), then S can either allow R to implement the decision or switch S off. The authors in [17] showed that the best action for R depends on R's uncertainty about S's utility and the rationality of S. If R is too certain about what S wants, and it knows S to be 'irrational', then it will have less incentive to defer to S (or, equivalently, to allow S to switch it off).

The analysis of the off-switch game by [17] was not fully developed within a formal game-theoretic framework. To address this gap, [36] reformulated the off-switch problem as a Bayesian game with incomplete information. In this formulation, Nature determines whether R is uncertain with probability p_u and whether S is rational with probability p_r. Although this is a more rigorous formulation of the off-switch problem from a game-theoretic perspective, it introduces an artificial setting. For example, in [36], a non-rational human is

[1]This is the normative moral theoretical perspective adopted in the works concerned in this article, see e.g. [24]. The fact that we are restricting to this perspective in our paper does not implies its endorsement (we may actually find such restriction problematic, but a philosophical analysis of the general problem of AI alignment is not the subject of this paper).

This work is licensed under CC BY 4.0

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:*

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

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

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

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:*

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

From the above lemma, we can prove:

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

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

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

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

and

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

and, therefore,

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

□

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*

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

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

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

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

$$p(\nu(x)|\nu(o)) = N\left(\nu(x); \mu_p(x) + \frac{K_p(x,o)}{K_p(o,o)}(\nu(o)-\mu_p(o)), K_p(x,x) - \frac{K_p^2(x,o)}{K_p(o,o)}\right). \quad (11)$$

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

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

Now observe that

$$E[m_1] = \int \left(\mu_p(x) + \frac{K_p(x,o)}{K_p(o,o)}(\nu(o) - \mu_p(o))\right) N(\nu(o); \mu_p(o), K_p(o,o))d\nu(o)d\nu(o) = \mu_p(x). \quad (13)$$

and

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

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

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

and so:

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

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

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

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

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

Note that

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

Therefore, we have that

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

□

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 $p\mu_p(x) + (1-p)\mu_p(o) + e - \gamma'|\mu_p(o)| \geq \max\left(\mu_p(x), \mu_p(o)\right)$. 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

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

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

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

We separate the sum:

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

The other term in the sum

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

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

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

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

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

Note that

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

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:

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

and

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

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.

modelled as an agent minimising utility, which is a counterintuitive assumption. Moreover, the formulation in [36] does not really model the AI-assistance game principles P1–P3, since R does not learn S's preferences.

In this paper, we revisit the off-switch problem and model it more correctly as a *signalling game*. Signalling games [15, 31] specifically refer to a class of two-player Bayesian games with incomplete information, where one player (S in our case) possesses private information, while the other (R in our case) does not. The informed player shares information with the uniformed one through a message. We make the assumption that, in the off-switch problem, the messages are S's preferences about some underlying decision problem. R uses these preferences to learn S's utilities and then chooses its optimal action.

In this setting, R's uncertainty arises statically from the task of learning from preferences. For S, we adopt the more realistic assumption that S is a *bounded rational* agent, that is an agent who behaves rationally within the limits of their cognitive abilities. We consider different *bounded rationality* mechanisms.

We reprove the results of [17] in this setting using real machine learning models to learn from preferences. We additionally show that S does not have any incentive to lie when sending a message to R. Furthermore, we also discuss how the cost of sending a message affects the optimal strategy in the game. Both [17] and [36] consider a single-utility scenario. In contrast, we also analyse scenarios involving incomparability. Such incomparability may arise due to bounded rationality or because S has multiple utility functions in mind.

2. Preliminary

We assume that the human S and the robot R aim to maximise S's preferences of an underlying decision problem. We consider a standard set up for decision-making [25] consisting of three key components: 1) a set $\mathcal{S}$ of finite states of the world; 2) a set $\mathcal{O}$ of outcomes; and 3) a set $\mathcal{X}$ of acts, which are mappings from states to outcomes. S expresses preferences over acts. We use the term preferences here in a more colloquial sense; it is, in fact, more realistic to assume that we can only observe choices over acts. Therefore, we can represent this decision making problem as a tuple $(\mathcal{S}, \mathcal{O}, \mathcal{X}, C)$, where $C : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{X})$ is a choice function ($\mathcal{P}(\mathcal{X})$ being the power-set of $\mathcal{X}$).

Example 2.1. Assume S's objective is to make a very good risotto. The outcome is the taste of the risotto, which is determined by its recipe. The decision-making problem is to find the best recipe. In this context, the state of the world can represent factors such as the quality of the ingredients, e.g., whether they are good or bad, which may significantly influence the resulting taste. Therefore, an act is represented by a vector specifying the recipe for each possible state of the world. If we assume that a recipe is a vector of $\mathbb{R}^c$, where c is the number of ingredients, then an act z is a vector in $\mathbb{R}^{2c}$, assuming only two states of the world good/bad. Then S expresses choices over the powerset of a finite subset of $\mathbb{R}^{2c}$ (that is, over subsets of recipes).

Hereafter, we also include examples of more standard instantiations of this problem, commonly considered in the foundations of decision-making under uncertainty. *Desirable gambles:* $\mathcal{S} = \Omega$ is a possibility space of an uncertain experiment (for instance, tossing a coin); $\mathcal{O} = \mathbb{R}$ and $\mathcal{X} = \mathbb{R}^{|\Omega|}$ is the set of all gambles [35]. *Anscombe-Aumann:* $\mathcal{S} = \Omega$; $\mathcal{O} = D(\mathcal{O})$ is the set of simple probability distributions on a finite set of prizes $\mathcal{O}$ and $\mathcal{X} \subset \mathbb{R}^{|\Omega|\times|\mathcal{O}|}$ [2]. A connection between these two settings is proven in [38, 39].

For a given decision problem $(\mathcal{S}, \mathcal{O}, \mathcal{X}, C)$, *rationality* of the decision-maker can be imposed by enforcing the choice function C to satisfy a set of constraints [1, 3, 28]. A minimal rationality requirement is usually:

$$C(A \cup B) = C(C(A) \cup B), \tag{1}$$

for all $A, B \subseteq \mathcal{P}(\mathcal{X})$, which is known as *path-independence* property [23]. Other constraints can be added depending on the specific decision-making problem [13, 27, 34] (including topological constraints). The choice mechanisms (procedures) are typically defined by specifying a rule that determines how to identify $C(A)$ for each input set A. A first example is the 'scalar optimisation choice' [1] defined as:

$$C(A) = \{z \in A : \text{there is no } y \in A \text{ s.t. } \nu(y) > \nu(z)\}, \tag{2}$$

that is, the choice function is defined by $\nu : \mathcal{X} \to \mathbb{R}$. We refer to ν as utility function. Another example is the 'vector optimisation choice' defined through the following Pareto-dominance criterion

$$C(A) = \{z \in A : \text{there is no } y \in A \text{ s.t. } \boldsymbol{\nu}(y) > \boldsymbol{\nu}(z)\}, \tag{3}$$

where $\boldsymbol{\nu} : \mathcal{X} \to \mathbb{R}^d$ is a vectorial function. Another choice function is provided below, selecting acts that are better with respect to at least one of the utilities:

$$C(A) = \left\{ \bigcup_{k=1,\dots,d} \arg\max_{z\in A} \nu_k(z) \right\}. \tag{4}$$

The definitions (3) and (4) coincide with *maximality* and *e-admissibility* when applied to gambles [12, 27]. Each one of these representations defines choice functions satisfying different 'rationality constraints'. For instance, all these three choice functions satisfy property (1).[2] We refer to the choice function in (2) as a binary preference.

[2] This holds when $\mathcal{X}$ is finite. In the infinite case, additional assumptions are needed.

Note that, here, we do not assume state-independence; therefore, the utility function $\boldsymbol{\nu}$ is generally state-dependent [26].

2.1. Learning from choices under rationality. The problem of learning from choices generalises the concept of learning from preferences. Given observations in the form of choices

$$\mathcal{D} = (A_i, C(A_i))_{i=1}^n, \tag{5}$$

and assuming the choice function satisfies rationality criteria such as (1), then it can be represented through a mechanism such as (2)–(4). In this case, learning a choice function reduces to the problem of learning a utility function vector [5, 9]. Since the unknown is a function we can use a Gaussian Process (GP) to learn $\boldsymbol{\nu}$ [5, 9]. In this section, we assume that for each $y, z \in \mathcal{X}$ then $\boldsymbol{\nu}(y) \neq \boldsymbol{\nu}(z)$. This assumption prevents issues with zero probability when calculating the posterior, as discussed in [5]. As a result, for instance, in the scalar optimisation case, the choice function defined in (2) is such that $C(A)$ contains only a single element for each set A. In the next section, we will generalise this setting by introducing a limit of discernibility.

To learn a choice function, we assume a GP prior on the latent unknown utility vector function:

$$\boldsymbol{\nu}(z) = \begin{bmatrix} \nu_1(z) \\ \nu_2(z) \\ \vdots \\ \nu_d(z) \end{bmatrix} \sim GP\left(\mu_0(z), K_0(z, z')\right), \tag{6}$$

where $\mu_0(z), K_0(z, z')$ are respectively, the prior mean and covariance functions (the parameters of the GP), and then use the data in (5) and the rationality constraints, which constrain $\boldsymbol{\nu}$ (as for instance in (4)) to compute a posterior over $\boldsymbol{\nu}$. The posterior is not a GP, but we can approximate it with a GP using various approaches (see below), leading to the posterior

$$\boldsymbol{\nu}(z) = \begin{bmatrix} \nu_1(z) \\ \nu_2(z) \\ \vdots \\ \nu_d(z) \end{bmatrix} \sim GP\left(\mu_p(z), K_p(z, z')\right), \tag{7}$$

where $\mu_p(z), K_p(z, z')$ are respectively, the posterior mean and covariance functions of the GP. Given this posterior, we can probabilistically predict the decision maker's choice for any new finite choice set B, that is we can compute $P(C(B) = B' | \mathcal{D})$ for each $B' \subseteq B$.

When $d = 1$ (scalar utility), the problem simplifies to the problem of learning a complete preference (aka standard preference learning).

Note that, for standard preference learning, the posterior is a SkewGP [6–8]. It means that the posterior is asymmetric (skewed). However, for the analytical derivation of the results in the paper, we will approximate it with a GP (which has Gaussian marginals). We can use three methods to compute this approximation: (i) Laplace's approximation [20, 37]; (ii) Expectation Propagation [21]; (iii) Kullback-Leibler divergence minimization [22], including Variational approximation [16] as a particular case.

Example 2.2. To introduce the problem of choice-function learning, we use a simple example: learning S's preferences over risotto recipes. For clarity, we consider a case where all ingredients are fixed except for the amount of butter with values in $[1, 9]$. We assume that in her mind, S's taste as a function of the butter amount x is as depicted in Figure 1, where higher values indicate a stronger preference. We further assume that there is only one state of the world, so that $c = 1$. R does not know S's taste and learn it indirectly from S's preferences, such as

$$\mathcal{D} = \{6.5 \succ 3.5, 7 \succ 5, 6.5 \succ 5.5, 3.5 \succ 8.5, \\ 1 \succ 9, 7 \succ 1.5, 4.5 \succ 7.5, 3.5 \succ 4\}$$

where $6.5 \succ 3.5$ means that S prefers a risotto with a butter amount of 6.5 over one with a butter amount of 3.5. These preferences have been generated according to the utility function depicted in Figure 1.

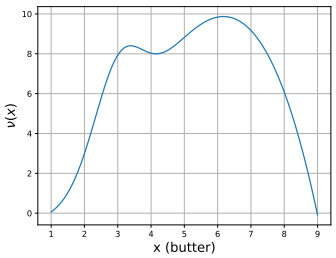


Figure 1. *S's utility for risotto.*

Since there is only a single latent utility dimension ($d = 1$), this problem reduces to learning a complete preference order. R will do it by placing a GP prior over the unknown utility $\nu(x)$, as depicted in Figure 2. The likelihood is simply:

$$p(\mathcal{D}|\nu) = I_{\{\nu(6.5)>\nu(3.5)\}} I_{\{\nu(7)>\nu(5)\}} \cdots I_{\{\nu(3.5)>\nu(4)\}} \tag{8}$$

where $I_{\{\nu(6.5)>\nu(3.5)\}}$ denotes the indicator function, which is equal to 1 when the condition in the subscript holds true, and 0 otherwise. The posterior computed based on the 8 preferences above is shown in Figure 3. The posterior computed based on a total of 30 preferences above is shown in Figure 4. It is important to note that R can never fully estimate the utility hidden in S's mind because this utility is not identifiable from preferences alone. R can only learn it up to a non-decreasing transformation. This explains why the posterior mean in Figure 4 is not 'exactly' equal to the original utility in Figure 2, but they both generate the same preferences.

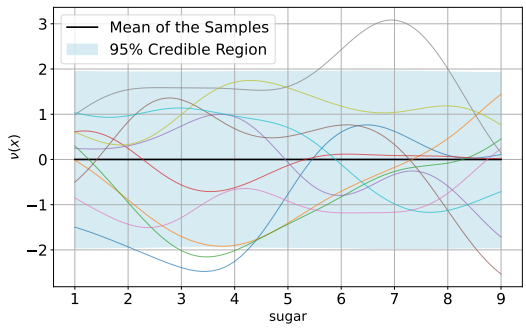


Figure 2. *GP prior: mean function (black line), 95% credible region (blue shaded area), and 10 samples of $\nu(x)$, each shown in a different colour.*

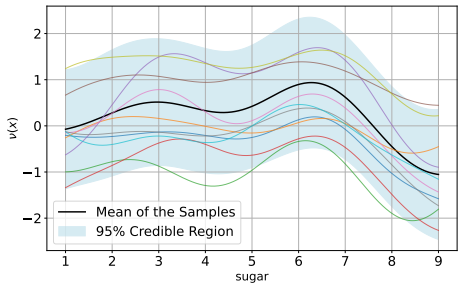


Figure 3. *GP posterior: mean function (black line), 95% credible region (blue shaded area), and 10 samples of $\nu(x)$, each shown in a different colour.*

2.2. Learning from choices under bounded-rationality. In standard decision theory, it is assumed that the decision maker is rational. However, due to cognitive limitations, we can generally only assume that the decision maker is bounded-rational [30]. A typical instance arises when the decision maker has limited time to make a decision (such as when playing chess), and, under time pressure, may resort to a random choice. Other cases involve limitations in computational resources [11, 14] and discernibility [19]. In the latter, the decision-maker may struggle to differentiate between two acts with similar utilities, resulting in a random choice. These situations are usually modelled through *random utility models* [33], where noise with a certain distribution is included in the choice mechanism. For instance, in this case, the decision maker is assumed to choose $C(\{y,z\}) = \{z\}$ when

$$\nu(z) + n(z) > \nu(y) + n(y) \tag{9}$$

where $n(z), n(y)$ are independent noises. If $n(z), n(y) \sim N(0,\sigma^2)$ [32], then

$$\begin{aligned} P(C(\{z,y\}) = \{z\}) &= \Phi\left(\frac{\nu(z)-\nu(y)}{\sigma}\right), \\ P(C(\{z,y\}) = \{y\}) &= \Phi\left(\frac{\nu(y)-\nu(z)}{\sigma}\right), \end{aligned} \tag{10}$$

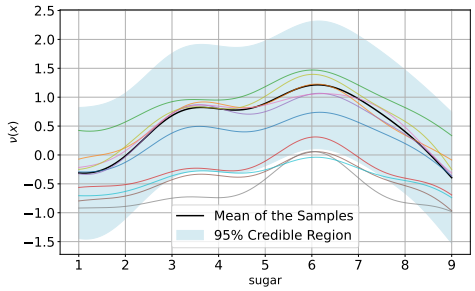


Figure 4. *GP posterior: mean function (black line), 95% credible region (blue shaded area), and 10 samples of $\nu(x)$, each shown in a different colour.*

where Φ is the CDF of the standard normal distribution. In the paper, we use ϕ to denote the PDF of the standard normal distribution. It can be noticed that when $\nu(z)-\nu(y) = 0$, the decision maker chooses between the two options with probability $1/2$. However, when the difference $|\nu(z)-\nu(y)|$ is very large compared to σ, the decision maker will choose deterministically. Therefore, this random utility model captures a limit of discernibility through the discernibility parameter σ [5]. Note that, we can use the GP model to learn from this noisy data, in this case the likelihood (8) becomes

$$p(\mathcal{D}|\nu) = \Phi\left(\tfrac{\nu(6.5)-\nu(3.5)}{\sigma}\right)\cdots\Phi\left(\tfrac{\nu(3.5)-\nu(4)}{\sigma}\right). \tag{11}$$

These models are used for problem where the decision-maker needs to make a single item choice. However, if we allows incompleteness then we can define choice functions modelling bounded rational agents. For instance, the following choice function models a limit of discernibility through incompleteness [19]:

$$C(\{z,y\}) = \begin{cases} \{z\} & \text{if } \nu(z) > \nu(y)+\sigma, \\ \{y\} & \text{if } \nu(y) > \nu(z)+\sigma, \\ \{z,y\} & \text{if } |\nu(z)-\nu(y)| \le \sigma, \end{cases} \tag{12}$$

where $\sigma > 0$ is the limit of discernibility. In this section, we provided examples of bounded rationality for the scalar optimisation choice mechanism. Similar models can also be introduced for the vector mechanism [5].

3. Modified Signalling Games

Signalling games [15, 31] specifically refer to a class of two-player games with incomplete information, where one player possesses private information (the Sender), while the other does not (the Receiver). The timing and payoffs of the game are as follows:

1. Nature draws a state of the world (a type) $t_i \in T$ according to a probability distribution $p(t)$.
2. The Sender observes t_i and then chooses a message $m_j \in M$ from a finite set of messages M and sends it to the Receiver.

3. The Receiver observes m_j (but not t_i) and then chooses an action $a_k \in A$ from a finite set of possible actions A.
4. If $a_k \in A' \subset A$ then the Sender chooses an action $b_l \in B$ from a finite set of possible actions B. Otherwise, $b_l = \varnothing$ (null decision).
5. Payoffs for Sender and Receiver are given by $u_S(t_i, m_j, a_k, b_l)$ and, respectively, $u_R(t_i, m_j, a_k, b_l)$.

Step 4 in the game is typically absent in standard signalling games, which is why we refer to it as *modified*.

Requirement 1. After observing any message m_j from M, the Receiver must have a belief about which types could have sent m. Denote this belief by the probability $p(t|m_j)$.

Requirement 2R. For each $m_j \in M$, the Receiver's action $a^*(m_j)$ must maximise the Receiver's expected utility, given the belief $p(t|m_j)$ about which types could have sent m_j. That is,

$$a^*(m_j) \in \arg\max_{a_k \in A} \int_T u_R(t, m_j, a_k, b^*(a_k)) dp(t|m_j). \tag{13}$$

Requirement 2R also applies to the Sender, but the Sender has complete information (and hence a trivial belief), so Requirement 2S is simply that the Sender's strategy be optimal given the Receiver's strategy:

Requirement 2S. For each $t_i \in T$, the Sender's message $m^*(t_i)$ and the action $b^*(a^*(m_j))$ must maximize the Sender's utility, given the Receiver's strategy $a^*(m_j)$. That is, we have that:

$$\begin{aligned} &\big(m^*(t_i), b^*(a^*(m_j))\big) \\ &\in \arg\max_{m \in M, b \in B} u_s(t_i, m, a^*(m), b(a^*(m))). \end{aligned} \tag{14}$$

Finally, given the Sender's strategy $m^*(t_i)$, let T_j denote the set of types that send the message m_j. That is, t_i is a member of the set T_j if $m^*(t_i) = m_j$. If T_j is nonempty then the information set corresponding to the message m_j is on the equilibrium path; otherwise, m_j is not sent by any type and so the corresponding information set is off the equilibrium path. For messages on the equilibrium path, we have that

Requirement 3. For each $m_j \in M$, if there exists $t \in T$ such that $m^*(t) = m_j$, then the Receiver's belief at the information set corresponding to m_j must follow from Bayes' rule and the Sender's strategy:

$$p(t|m_j) = \frac{I_{\{m^*(t)=m_j\}}(t) p(t)}{\int_T I_{\{m^*(t)=m_j\}}(t) dp(t)}. \tag{15}$$

Definition 3.1. A pure-strategy perfect Bayesian equilibrium in a signalling game is a triplet of strategies $m^*(t_i), a^*(m_j), b(a^*(m_j))$ and a belief $p(t|m_j)$ satisfying Requirements 1–3.

4. The off-switch game as a signalling game

In this section, we revisit the off-switch problem [17] through the lens of signalling games. In the off-switch game, the human (S), acts as the Sender and the robot (R), as the Receiver, with T representing the set of possible types for S. Each type $t_i \in T$ characterises S's "taste" and degree of bounded rationality for an underlying decision problem $(\mathcal{S}, \mathcal{O}, \mathcal{X}, C)$. As studied in [17], we consider a bounded-rationality model similar to (9), where a single scalar utility is assumed. Thus, the type t_i is a realisation of a utility function ν and (a vector of) noises[3] $\mathbf{n}$, where $p(\mathbf{n}) = N(\mathbf{n}; 0, I\sigma^2)$ and $p(\nu) = GP(\nu; \mu_0, K_0)$, for some mean function μ_0, kernel function K_0, and parameter σ^2 (I is the identity matrix). We further assume that all these parameters are common knowledge in the game.

The message $m \in M$ encapsulates S's preferences/choices over acts in $\mathcal{X}$, relative to the underlying decision problem $(\mathcal{S}, \mathcal{O}, \mathcal{X}, C)$. That is, m represents a dataset of choices as in (5). Since we are considering a single utility function, we assume a choice mechanism as in (2). We first assume that the payoffs do not directly depend on the message m. In signalling games, this assumption is referred to as *cheap-talk* [15]. Furthermore, we assume that each type t can only choose one message, i.e., S can only send the message determined by Nature through S's type. This is the setting studied in [17]. The message is a choice dataset, such as the one in (5), which R uses to estimate S's utility using the approach described in Section 2.1.

We consider two acts $x, o \in \mathcal{X}$ relative to the underling decision problem $(\mathcal{S}, \mathcal{O}, \mathcal{X}, C)$. We assume that o is the status-quo and x is some new act. For instance, in our risotto example, o could be the best recipe known by S and x a new recipe proposed by R.

Remark 4.1. Although we do not address the optimal selection of x in this paper, we assume that R will choose an act x that is preferable to o whenever possible. This assumption explains why S is willing to engage in the game with R: S expects R to recommend actions that improve upon the status quo, since S does not know how to identify such actions on their own. However, the choice of x may not be optimal when R is uncertain about the utility of S. In machine learning, the optimal selection of x can be done through a technique called Bayesian optimisation.

In this context, the available actions for R are $A = \{IMM, DEF, DoN\}$. IMM means 'immediate decision', that is R will implement x. DoN means 'do nothing'. DEF means R will defer to S. In the signalling game, this implies that $A' = \{DEF\}$. In the case R defers to S, then S has two possible actions, $B = \{OFF, \neg OFF\}$, either to

[3] It is a vector because the noise is also present in the message, that is in the choice dataset.

switch off R or to allow R to implement x. Therefore, the payoff for R is

$$\begin{aligned} u_R(t, IMM, \varnothing) &= \nu(x), \\ u_R(t, DoN, \varnothing) &= \nu(o), \\ u_R(t, DEF, b^*(DEF)) &= \nu(o) I_{\{\nu(o)+n(o)>\nu(x)+n(x)\}} \\ &\quad + \nu(x) I_{\{\nu(x)+n(x)>\nu(o)+n(o)\}}. \end{aligned} \tag{16}$$

Indeed, if R chooses IMM, then the payoff for R is equal to the utility that S receives from the act x. If R chooses DoN, then the payoff for R is equal to the utility of the status-quo for S (utility relative to the act o). The last case represents the payoff for the action DEF, where the action depends on S's move in the game, which is encapsulated by the indicator functions. Given S is a bounded rational agent, S chooses the action OFF or $\neg OFF$ based on S's noisy utility. If S chooses OFF, then S receives the utility of the status-quo o, otherwise the utility of x.

According to **Requirement 2R**, R will choose the action that maximises the expected value of the payoff in (16), where the expectation is computed with respect to $p(t|m_j)$, i.e., the posterior distribution over S's type (utility ν for the underlying decision problem $(\mathcal{S}, \mathcal{O}, \mathcal{X}, C)$) learned by R from the message m_j (which is a dataset of preferences). We can then prove the following lemma.

Lemma 4.1. *Assume that $p(\nu|\mathcal{D}) = GP\left(\nu; \mu_p, K_p\right)$ is the GP posterior computed by R from the prior $p(\nu) = GP\left(\nu; \mu_0, K_0\right)$, the bounded-rationality likelihood (10) and the message $m_j = \mathcal{D}$, then the expected payoffs of R's actions are:*

$$\begin{aligned} &DEF: \int_T \left(\nu(o) I_{\{\nu(o)+n(o)>\nu(x)+n(x)\}} + \nu(x) I_{\{\nu(o)+n(o)<\nu(x)+n(x)\}}\right) \\ &dp(\nu(x), n(x), \nu(o), n(o)|m_j) \\ &= \mu_p(x)\left(1 - \Phi\left(\frac{\mu_p(o)-\mu_p(x)}{\sqrt{K_p(x,x)+2\sigma^2+K_p(o,o)-2K_p(x,o)}}\right)\right) \\ &+ \frac{K_p(x,x)-K_p(x,o)}{\sqrt{K_p(x,x)+2\sigma^2+K_p(o,o)-2K_p(x,o)}} \phi\left(\frac{\mu_p(o)-\mu_p(x)}{\sqrt{K_p(x,x)+2\sigma^2+K_p(o,o)-2K_p(x,o)}}\right) \\ &+ \mu_p(o)\left(1 - \Phi\left(\frac{\mu_p(x)-\mu_p(o)}{\sqrt{K_p(x,x)+2\sigma^2+K_p(o,o)-2K_p(x,o)}}\right)\right) \\ &+ \frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(x,x)+2\sigma^2+K_p(o,o)-2K_p(x,o)}} \phi\left(\frac{\mu_p(x)-\mu_p(o)}{\sqrt{K_p(x,x)+2\sigma^2+K_p(o,o)-2K_p(x,o)}}\right), \end{aligned} \tag{17}$$

$$IMM: \int_T \nu(x) dp(\nu(x), n(x), \nu(o), n(o)|m_j) = \mu_p(x), \tag{18}$$

$$DoN: \int_T \nu(o) dp(\nu(x), n(x), \nu(o), n(o)|m_j) = \mu_p(o), \tag{19}$$

where

$$p(\nu(x), n(x), \nu(o), n(o)|m_j) = p(\nu(x), \nu(o)|m_j) p(n(x), n(o))$$

with $p(n(x), n(o)) = N(n(x); 0, \sigma^2) N(n(o); 0, \sigma^2)$ *and*

$$p(\nu(x), \nu(o)|m_j) = N\left(\begin{bmatrix}\nu(x)\\ \nu(o)\end{bmatrix}; \begin{bmatrix}\mu_p(x)\\ \mu_p(o)\end{bmatrix}, \begin{bmatrix}K_p(x,x) & K_p(x,o)\\ K_p(o,x) & K_p(o,o)\end{bmatrix}\right). \tag{20}$$

Proofs of this and the following results are provided in the Supplementary Material. In Lemma 4.1, we have exploited the fact that the marginal of a GP is a multivariate normal (equation (20)). We then introduce the following definitions:

Definition 4.1. We say that:

- S is **rational** whenever $\sigma \to 0$; S is **bounded-rational** otherwise;
- R has **no uncertainty** on S's utility whenever $K_o(x,x), K_o(o,o), K_o(o,x) \to 0$ (that is the prior becomes a Dirac's delta); otherwise R has **uncertainty**.

In the signalling game defined in Section 3, note that R having no uncertainty about S's utility implies that R possesses perfect knowledge of S's utility. This is because the prior $p(t)$ is common knowledge within the game. We consider

Proposition 4.1. *The optimal decisions for R are:*

- *If S is **rational** and R has **no uncertainty**, then DEF is never dominated by IMM, DoN.*
- *If S is **bounded-rational** and R has **no uncertainty**, then DEF is never optimal;*
- *If S is **rational** and R has **uncertainty**, then DEF is always optimal.*
- *If S is **bounded-rational** and R has **uncertainty**, then DEF is optimal if (17) is larger or equal than the maximum between (18) and (19).*

Therefore, by interpreting the non-optimality of DEF as the robot disabling the off-switch,[4] we conclude that:

> If S is rational, then R will never disable the off-switch. If S is bounded-rational, a necessary condition for R not to disable the off switch is the presence of uncertainty.

This result aligns with what was proven in [17]; however, in our case, the statements have been rigorously established within the framework of signalling games, also incorporating the posterior uncertainty (in equation (20)) derived from preference learning. This means that the result was proven under the principles P1–P3.

[4] This is based on the premise that the robot (as a recommendation system) will propose an act x that improves S's utility whenever possible (compared to o). In this scenario, deferring to a human is equivalent to not disabling the off-switch since, otherwise, R will implement x.

It is particularly interesting to note that the conclusions of Proposition 4.1 would still hold, even if the absence of uncertainty did not imply perfect knowledge of ν for the robot.[5] For this reason, the *moral* of this result is that we should not build R to estimate ν through a model, such as a *neural network*, that does not provide any measure of uncertainty [17]. Indeed, Proposition 4.1 strongly supports the use of probabilistic methods in AI. We will revisit this in Section 6.1.

4.1. The cost of messaging. In the previous section, we have considered a situation where messaging is 'cheap', that is it does not affect S's utility. If this is not the case, we may assume that the cost of communications is proportional to the length of the message m, that is, the communication cost is $\gamma'\ell_m$ for some scaling parameter $\gamma' > 0$. However, since the utility ν is only defined by pairwise comparisons and lacks an absolute scale, we cannot directly assign a value to γ'. Instead, the communication cost must be defined relative to ν, for example, as $\gamma' = \gamma|\nu(o)|$ for some $\gamma > 0$. Using our risotto example, this implies that the communication cost is expressed on a "taste" scale. In this case, the payoff for R is:

$$\begin{aligned} u_R(t, DEF, b^*(DEF)) &= \nu(o) I_{\{\nu(o)+n(o)>\nu(x)+n(x)\}} \\ &\quad + \nu(x) I_{\{\nu(x)+n(x)>\nu(o)+n(o)\}} \\ &\quad - \gamma(\ell_{m_j} + 1), \\ u_R(t, IMM, \varnothing) &= \nu(x) - \gamma\ell_{m_j}, \\ u_R(t, OFF, \varnothing) &= \nu(o) - \gamma\ell_{m_j}. \end{aligned} \tag{21}$$

Here, $\gamma\ell_{m_j}$ represents the communication cost associated with sending the message m_j, and the additional γ in the first term arises because deferring a decision to S involves further communication.

Corollary 4.1. *The optimal decisions for R are:*

- *If R has* ***no uncertainty****, then DEF is never optimal.*
- *If S is* ***rational*** *and R has* ***uncertainty****, then DEF is optimal if*

$$p\mu_p(x) + (1-p)\mu_p(o) + e \geq \beta + \max\left(\mu_p(x), \mu_p(o)\right), \tag{22}$$

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*

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

- *If S is* ***bounded-rational*** *and R has* ***uncertainty****, then DEF is optimal if* (17) *is larger or equal than* β *plus the maximum between* (18) *and* (19)*.*

[5] The result of the Proposition depends only on whether the posterior is deterministic or not.

where $\beta = \gamma\mu_p(o)\left(1 - 2\Phi\left(\frac{-\mu_p(o)}{\sqrt{K_p(o,o)}}\right)\right) + 2\gamma\sqrt{K_p(o,o)}\phi\left(\frac{-\mu_p(o)}{\sqrt{K_p(o,o)}}\right)$.

Summarising it we have that

> With communication cost, if R has no uncertainty about S's utilities, then R will always disable the off switch (even when S is rational).

The takeaway is that if S genuinely values preserving the off switch, it should not penalise messaging to R.

4.2. Another mechanism of bounded rationality. In the previous two sections, we considered a utility model governed by a Gaussian-noise bounded rationality mechanism. In this section, we analyse the bounded rationality mechanism described in (12), which captures incomparability arising from a limit of discernibility. In this case, we need to define a payoff for DEF when $|\nu(x) - \nu(o)| \leq \sigma$, that is when S cannot distinguish between the two acts x, o. In this case, we assume that the payoff is represented as the set $\{\nu(x) - \epsilon; \nu(o) - \epsilon\}$. Indeed, since S cannot distinguish the two acts and chooses both of them, S's payoff is a set, $\{\nu(x); \nu(o)\}$, and $\epsilon \in (0, \sigma]$ is a penalisation term introduced to penalise S's payoff for being imprecise. This penalisation term is similar to the one introduced in [40] for evaluating imprecise classifiers. Using this framework, we can prove the following lemma.

Lemma 4.2. *Assume that* $p(\nu|\mathcal{D}) = GP\left(\nu; \mu_p, K_p\right)$ *is the GP posterior computed by R from the prior* $p(\nu) = GP\left(\nu; \mu_0, K_0\right)$*, the likelihood* (12) *and the message* $m_j = \mathcal{D}$*, then the expected payoffs of R's actions are:*

$$\begin{aligned} DEF : &\int_T \Big(\nu(o) I_{\{\nu(o)>\nu(x)+\sigma\}} + \nu(x) I_{\{\nu(x)>\nu(o)+\sigma\}} \\ &+ \{\nu(x), \nu(o)\} I_{|\nu(o)-\nu(x)|\leq\sigma}\Big) dp(\nu(x), \nu(o)|m_j) \\ &= \mu_p(x)\left(1 - \Phi\left(\frac{(\mu_p(o)+\sigma-\mu_p(x))}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right)\right) \\ &+ \frac{K_p(x,x)-K_p(x,o)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\phi\left(\frac{\mu_p(o)+\sigma-\mu_p(x)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right) \\ &+ \mu_p(o)\left(1 - \Phi\left(\frac{(\mu_p(x)+\sigma-\mu_p(o))}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right)\right) \\ &+ \frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\phi\left(\frac{\mu_p(x)+\sigma-\mu_p(o)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right) \\ &+ \Bigg\{\mu_p(o) - \mu_p(o)\Bigg(2 - \Phi\left(\frac{(\mu_p(x)+\sigma-\mu_p(o))}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right) \\ &- \Phi\left(\frac{(-\mu_p(x)+\sigma+\mu_p(o))}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right)\Bigg) \\ &+ \frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\left(\phi\left(\frac{-\mu_p(x)+\sigma+\mu_p(o)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right)\right. \\ &\left.- \phi\left(\frac{\mu_p(x)+\sigma-\mu_p(o)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right)\right) - \epsilon; \end{aligned}$$

$$\mu_p(x) - \mu_p(x)\Bigg(2 - \Phi\left(\frac{(\mu_p(x)+\sigma-\mu_p(o))}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right) - \Phi\left(\frac{(-\mu_p(x)+\sigma+\mu_p(o))}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right)\Bigg) + \frac{K_p(o,o)-K_p(x,o)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\Bigg(\phi\left(\frac{-\mu_p(o)+\sigma+\mu_p(x)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right) - \phi\left(\frac{\mu_p(o)+\sigma-\mu_p(x)}{\sqrt{K_p(x,x)+K_p(o,o)-2K_p(x,o)}}\right)\Bigg) - \epsilon\Bigg\}, \tag{23}$$

$$IMM : \int_T \nu(x) dp(\nu(x)n(o)|m_j) = \mu_p(x), \tag{24}$$

$$DoN : \int_T \nu(o) dp(\nu(x), \nu(o)|m_j) = \mu_p(o), \tag{25}$$

where $p(\nu(x), \nu(o)|m_j)$ is defined in (20).

The payoff for DEF can be a set and this leads to incomparability, a complete preference order cannot be established. Therefore, the players are restricted to making comparisons based solely on dominance. We consider the two possible dominance conditions defined in (3) (we refer to it as criterion (A)) and (4) (we refer to it as criterion (B)).

Proposition 4.2. *The optimal decisions for R are:*

- *If S is **rational** and R has **no uncertainty**, then DEF is never dominated.*
- *If S is **bounded-rational** and R has **no uncertainty**, DEF is always dominated whenever $|\nu(x)-\nu(o)| \leq \sigma$, otherwise DEF is not dominated.*
- *If S is **rational** and R has **uncertainty**, then DEF is always optimal.*
- *If S is **bounded-rational** and R has **uncertainty**, then DEF is optimal if*
  - **(A)** *the minimum of (23) is larger or equal than the maximum between (24) and (25).*
  - **(B)** *the maximum of (23) is larger or equal than the maximum between (24) and (25).*

We can then conclude that:

> If S is rational, then R will never disable the off-switch. If S is bounded-rational, a necessary condition for R not to disable the off switch is the presence of uncertainty.

These conclusions are similar to those derived from Proposition 4.1 but, in this case, we have proved them with a mechanism of bounded rationality that is fully deterministic.

4.3. Lies and deception. Consider a choice set comprising of two acts $A = \{y, z\}$, then, according to the bounded-rationality model in the previous section, S can send three possibles messages:

$$\text{either} \quad C(A) = \{y\} \quad \text{or} \quad C(A) = \{z\} \quad \text{or} \quad C(A) = \{y, z\}.$$

This is true for any A_i in $\mathcal{D} = (A_i, C(A_i))_{i=1}^n$ and, therefore, there are 3^n possible messages in this game. In the previous section, we assumed that S sends the message defined by their type, that is, for instance, $C(A) = \{y\}$ if $\nu(y) > \nu(z) + \sigma$. We call this message the *honest message*.

Since in the AI-assistance game, R aims to maximise S's payoffs. It is easy to verify the following:

Proposition 4.3. *In the AI-assistance game, sending the honest message is always the best action for S,*

or, in other words,

> In the AI-assistance game, S does not have incentives to deceive R.

For a bounded-rationality model like the one in (9) (with random noise), S may be in a situation where $\nu(x) > \nu(o)$, but $\nu(x) + n(x) < \nu(o) + n(o)$ (due to noise, and thus due to S's bounded rationality). In other words, if deferred to, S would ultimately select o based on their noisy preferences. In this case, if S knows their noisy preference in advance, the best action for S is to send a message to R that results in R computing the wrong estimate, $\mu_p(x) < \mu_p(o)$. In this respect, R would also consider o to be better than x.

By maximising its own payoff, R would also maximise S's payoff. However, this strategy depends on the realisation of the noise and would not be the optimal strategy in the context of repeated games, where the two players interact after sending the message at the beginning of the game. In this case, S would aim to maximise their expected payoff in the long run. We will leave the proof of this to future work, as it involves consistency results for repeated games. For the remainder of the paper, we will assume that S will always send the *honest message*.

5. Vector-valued payoffs

In the previous sections, we have assumed the underlying decision problem to have a single utility. In a case where we have competing utilities, this leads to vector-valued payoff [29] in the AI-assistance game. In this case, the payoff in (16) becomes:

$$\begin{aligned} u_R(t, DEF, b^*(DEF)) &= \boldsymbol{\nu}(o) I_{\{\boldsymbol{\nu}(o)+\mathbf{n}(o) \succ \boldsymbol{\nu}(x)+\mathbf{n}(x)\}} \\ &\quad + \boldsymbol{\nu}(x) I_{\{\boldsymbol{\nu}(x)+\mathbf{n}(x) \succ \boldsymbol{\nu}(o)+\mathbf{n}(o)\}}, \\ u_R(t, IMM, \varnothing) &= \boldsymbol{\nu}(x), \\ u_R(t, DoN, \varnothing) &= \boldsymbol{\nu}(o). \end{aligned} \tag{26}$$

Therefore, a complete preference order cannot be established due to potentially conflicting utilities. Players are

restricted to making comparisons based solely on dominance. We consider the dominance condition $\succ$ as discussed earlier in (3) (Pareto dominance) or in (4), along with the random noise model for bounded rationality.

Proposition 5.1. *The optimal decisions for R are:*

- *If S is **rational** and R has **no uncertainty**, then DEF is never dominated by IMM,DoN.*
- *If S is **bounded-rational** and R has **no uncertainty**, then DEF is always dominated.*
- *If S is **rational** and R has **uncertainty**, then DEF is never dominated.*
- *If S is **bounded-rational** and R has **uncertainty**, the optimality of DEF depends on the specific case.*

The results in Proposition 5.1 are practically the same as in Proposition 4.1. The reason is that we have stated the results only for the dominance case, that is when an act is better than the other act. In all other cases, R would be undecided because actions are incomparable. We can expect is that, since estimating a vector of utilities is more difficult than estimating one utility, higher uncertainty would lead R to defer more often to S.

6. Numeric experiments

6.1. Bounded rationality of R. In the previous sections, we have always assumed that R is fully rational. However, even R will have limited rationality, at least due to limited computational resources. Here, we will numerically assess the effect of R's rationality. We will do this by comparing the following methods of approximating the posterior distribution $p(t|m_j)$ (computing the posterior is the computational bottleneck for R). The distribution $p(t|m_j)$ is the posterior of ν given the choice-data. We consider the following approximations of the posterior:

- MAP: R computes the maximum a-posteriori estimate for ν. This means there is no uncertainty representation. This is equivalent to estimating the utility using a Neural Network (NN) model.
- Laplace: The Laplace approximation is used to approximate the posterior with a GP, where the mean is the MAP estimate for ν and the covariance is equal to the observed Fisher information matrix.
- EP: Expectation Propagation, which approximates the posterior with a GP whose mean and covariance are computed through moment matching.
- SkewGP: The posterior is a SkewGP, meaning there is no Gaussian approximation. Instead, sampling is necessary to compute inferences.

In terms of computational load, for a small dataset of preferences, the order from the cheapest to the heaviest is NN, Laplace, EP, and SkewGP.

We performed 1000 Monte Carlo simulations in which a utility ν was sampled from a GP with zero mean and a square-exponential kernel, with randomly generated length-scale and variance. We used the generated ν to create a dataset of 30 noisy preferences (with $\sigma = 1$), which represents the message. We then simulated R computing the posterior using the four approximations discussed above. For each case, we computed the optimal decision (DoN, IMM, DEF) for R.

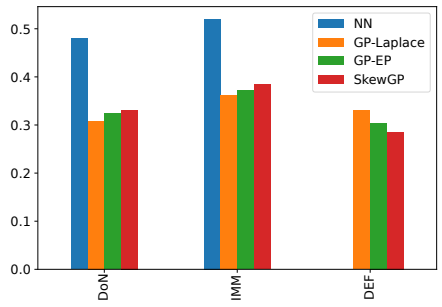


Figure 5. *Percentage of decisions for the four approximations, with MAP denoted as NN.*

Figure 5 reports the percentage of decisions for the three actions. As proven in Proposition 4.1, a robot R that does not model uncertainty never defers to the human. This results is confirmed here as for the NN (MAP) based inference DEF is never optimal. For the other posterior approximations, the differences in the decisions are relatively small. SkewGP provides the decision closest to the optimal one. It is also well-known that EP provides a better approximation of the posterior than Laplace. However, the key message is that it is better to be approximately Bayesian (rational) than to ignore uncertainty entirely.

7. Conclusion and future works

In this paper, we have formulated the off-switch problem as a signalling game between a robot and a human. This approach allows us to model the problem as it would be implemented in practice through a machine learning framework. The human communicates their preference to the robot, which then uses this information to estimate the human's utilities (through some machine learning model) and decide whether to do nothing, take immediate action, or defer to the human. If the robot chooses not to defer, we interpret this as disabling its off-switch.

In this more realistic setting, as demonstrated in the original paper [17], we have proven that if the human is fully rational, the robot will never disable the off-switch. However, if the human is bounded-rational, a necessary condition for the robot to avoid disabling the off-switch is the presence of uncertainty. We have validated this statement under various models of bounded rationality, including scenarios with vector-valued payoffs. The key takeaway is that AI systems should not be built using

machine learning models that fail to account for uncertainty—such models risk disabling their own off-switch.

As future works, we would like to investigate how these results depend on the choice of the prior. In our example, we assumed that the prior is common knowledge in the game, but this is unlikely to hold in real-world applications. A key question is whether we can leverage knowledge of human utility to design priors that inherently favour deferring to humans. This is particularly important for high-stakes applications. For instance, approaches similar to those in [4, 10] could be explored to achieve this.

Finally, the exchange scenario between an AI system and a human at the core of this work may actually involve more than one interaction step, meaning that it may be best interpreted (modelled) as a *repeated* signalling game [18]. What are the consequences of framing the off-switch problem in such a dynamic setting is yet another question that we are planing to tackle next.

References

[1] Mark Aizerman and Andrew Malishevski. “General theory of best variants choice: Some aspects”. In: *IEEE Transactions on Automatic Control* 26.5 (1981), pp. 1030–1040.

[2] F. J. Anscombe, R. J. Aumann, et al. “A definition of subjective probability”. In: *Annals of Mathematical Statistics* 34.1 (1963), pp. 199–205.

[3] Kenneth J Arrow. “Rational choice functions and orderings”. In: *Economica* 26.102 (1959), pp. 121–127.

[4] A Benavoli, Luigi Chisci, A Farina, L Ortenzi, and G Zappa. “Hard-constrained versus soft-constrained parameter estimation”. In: *IEEE Transactions on aerospace and electronic systems* 42.4 (2006), pp. 1224–1239.

[5] Alessio Benavoli and Dario Azzimonti. *A tutorial on learning from preferences and choices with Gaussian Processes*. 2024. arXiv: 2403.11782.

[6] Alessio Benavoli, Dario Azzimonti, and Dario Piga. “Skew Gaussian processes for classification”. In: *Machine Learning* 109.9 (2020), pp. 1877–1902.

[7] Alessio Benavoli, Dario Azzimonti, and Dario Piga. “A unified framework for closed-form nonparametric regression, classification, preference and mixed problems with Skew Gaussian Processes”. In: *Machine Learning* (Sept. 13, 2021), pp. 1–39.

[8] Alessio Benavoli, Dario Azzimonti, and Dario Piga. “Preferential Bayesian optimisation with Skew Gaussian Processes”. In: *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. 2021, pp. 1842–1850.

[9] Alessio Benavoli, Dario Azzimonti, and Dario Piga. “Learning choice functions with Gaussian processes”. In: *Uncertainty in Artificial Intelligence*. PMLR. 2023, pp. 141–151.

[10] Alessio Benavoli, Luigi Chisci, and Alfonso Farina. “Estimation of constrained parameters with guaranteed MSE improvement”. In: *IEEE transactions on signal processing* 55.4 (2007), pp. 1264–1274.

[11] Alessio Benavoli, Alessandro Facchini, Dario Piga, and Marco Zaffalon. “Sum-of-squares for bounded rationality”. In: *International Journal of Approximate Reasoning* 105 (2019), pp. 130–152.

[12] Jasper De Bock. “Archimedean choice functions: an axiomatic foundation for imprecise decision making”. In: *Information Processing and Management of Uncertainty in Knowledge-Based Systems: 18th International Conference, IPMU 2020, Lisbon, Portugal, June 15–19, 2020, Proceedings, Part II 18*. Springer. 2020, pp. 195–209.

[13] Jasper De Bock and Gert De Cooman. “Interpreting, axiomatising and representing coherent choice functions in terms of desirability”. In: *International Symposium on Imprecise Probabilities: Theories and Applications*. PMLR. 2019, pp. 125–134.

[14] Samuel J Gershman, Eric J Horvitz, and Joshua B Tenenbaum. “Computational rationality: A converging paradigm for intelligence in brains, minds, and machines”. In: *Science* 349.6245 (2015), pp. 273–278.

[15] Robert Gibbons. *Game theory for applied economists*. Princeton University Press, 1992.

[16] Mark N Gibbs and David JC MacKay. “Variational Gaussian process classifiers”. In: *IEEE Transactions on Neural Networks* 11.6 (2000), pp. 1458–1464.

[17] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. “The off-switch game”. In: *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

[18] Ayça Kaya. “Repeated signaling games”. In: *Games and economic behavior* 66.2 (2009), pp. 841–854.

[19] R Duncan Luce. “Semiorders and a theory of utility discrimination”. In: *Econometrica, Journal of the Econometric Society* (1956), pp. 178–191.

[20] David JC MacKay. “Bayesian methods for backpropagation networks”. In: *Models of neural networks III*. Springer, 1996, pp. 211–254.

[21] Thomas Peter Minka. “A family of algorithms for approximate Bayesian inference”. PhD thesis. Massachusetts Institute of Technology, 2001.

[22] Manfred Opper and Cédric Archambeau. “The variational Gaussian approximation revisited”. In: *Neural computation* 21.3 (2009), pp. 786–792.

[23] Charles R Plott. “Path independence, rationality, and social choice”. In: *Econometrica: Journal of the Econometric Society* (1973), pp. 1075–1091.

[24] Stuart Russell. *Human compatible: AI and the problem of control*. Penguin Uk, 2019.

[25] Leonard J Savage. *The foundations of statistics*. Courier Corporation, 1972.

[26] Mark J Schervish, Teddy Seidenfeld, and Joseph B Kadane. “State-dependent utilities”. In: *Journal of the American Statistical Association* 85.411 (1990), pp. 840–847.

[27] Teddy Seidenfeld, Mark J Schervish, and Joseph B Kadane. “Coherent choice functions under uncertainty”. In: *Synthese* 172 (2010), pp. 157–176.

[28] Amartya Sen. “The formulation of rational choice”. In: *The American Economic Review* 84.2 (1994), pp. 385–390.

[29] Lloyd S Shapley and Fred D Rigby. “Equilibrium points in games with vector payoffs”. In: *Naval Research Logistics Quarterly* 6.1 (1959), pp. 57–61.

[30] Herbert A Simon. “Bounded rationality”. In: *Utility and probability* (1990), pp. 15–18.

[31] M Spence. “Market Signaling”. In: *Harvard Univ. Press, Cambridge, MA* (1974).

[32] L.L. Thurstone. “A law of comparative judgment.” In: *Psychological review* 34.4 (1927), p. 273.

[33] Kenneth E Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

[34] Arthur Van Camp, Gert De Cooman, Enrique Miranda, and Erik Quaeghebeur. “Coherent choice functions, desirability and indifference”. In: *Fuzzy sets and systems* 341 (2018), pp. 1–36.

[35] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1991.

[36] Tobias Wängberg, Mikael Böörs, Elliot Catt, Tom Everitt, and Marcus Hutter. “A game-theoretic analysis of the off-switch game”. In: *Artificial General Intelligence: 10th International Conference, AGI 2017, Melbourne, VIC, Australia, August 15-18, 2017, Proceedings 10*. Springer. 2017, pp. 167–177.

[37] Christopher KI Williams and David Barber. “Bayesian classification with Gaussian processes”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.12 (1998), pp. 1342–1351.

[38] M. Zaffalon and E. Miranda. “Axiomatising incomplete preferences through sets of desirable gambles”. In: *Journal of Artificial Intelligence Research* 60 (2017), pp. 1057–1126.

[39] M. Zaffalon and E. Miranda. “Desirability foundations of robust rational decision making”. In: *Synthese* 198.27 (2021), pp. 6529–6570.

[40] Marco Zaffalon, Giorgio Corani, and Denis Mauá. “Evaluating credal classifiers by utility-discounted predictive accuracy”. In: *International Journal of Approximate Reasoning* 53.8 (2012), pp. 1282–1301.