Experiments with the User's Feedback in Preference Elicitation

Tereza Siváková[1], Miroslav Kárný[1]

[1]*The Czech Academy of Sciences, Institute of Information Theory and Automation*
182 00 Prague 8, Czech Republic

Abstract

This paper deals with user's preferences (wishes). Common users are uneducated in the decision-making (DM) theory and present their preferences incompletely. That is why we elicit them from such a user during the DM. The paper works with the DM theory called fully probabilistic design (FPD). FPD models closed DM loop, made by the user and the system, by the joint probability density (pd, real pd). A joint ideal pd quantifies the user's preferences. It assigns high probability values to preferred closed-loop behaviors and low values to undesired behaviors. The real pd should be kept near the ideal pd. By minimizing the Kullback-Leibler divergence of the real and ideal pds, the optimal decision policy is found. The presented algorithmic quantification of preferences provides ambitious but potentially reachable DM aims. It suppresses demands on tuning preference-expressing parameters. The considered ideal pd assigns high probabilities to desired (ideal) sets of states and actions. The parameters of the ideal pd (tuned during the DM via the user's feedback) are: ▶ relative significance of respective probabilities; ▶ a parameter balancing exploration with exploitation. Their systematic tuning solves meta-DM level task, which observes the agent's satisfaction expressed humanly by "school-marks". It opts free parameters to reach the best marks. A formalization and solution of this meta-task were recently done, but experience with it is limited. This paper recalls the theory and provides representative samples of extensive up to now missing simulations.

Keywords

Preference elicitation, Adaptive agent, Decision making, Bayes rule

Motivation Our results contribute to long-term research that tries to create a normative theory of dynamic decision making applicable by imperfect decision makers, [12, 13], [14, 16]. Its aims are close to the quest for universal artificial intelligence, [9, 15, 21].

1. Introduction

Decision making (DM) is the everyday activity of every human. It is important to make the right decisions to achieve the goal. DM is described by a closed-loop formed by an agent (the person, who makes decisions) and an environment. The environment of the agent is usually called a system and its dynamics is unknown. It is described by transition probability density (pd) between its states conditioned by the agent's actions. The agent observes the state s of the system and makes an action a to meet their wishes, ideally, to move the system to the desired state. The actions are chosen via the agent's policy π. It consists of decision rules r, which determine what

Italian Workshop on Artificial Intelligence and Applications for Business and Industries 2022, Udine, Italy

sivakter@cvut.cz (T. Siváková); school@utia.cas.cz (M. Kárný)
https://www.utia.cas.cz/AS (M. Kárný)
0000-0002-7440-6041 (M. Kárný)

action should be chosen in each time epoch depending on the system's state and the model of the system. The model m expresses the agent's beliefs about the dynamics of the real system.

The main task of DM is to select the optimal policy. This paper uses a fully probabilistic design (FPD), which introduces an ideal probability density

$$\mathsf{c}^i(b) = \prod_{t\in\mathbb{T}} \mathsf{m}^i(s_t|a_t, s_{t-1})\mathsf{r}^i(a_t|s_{t-1}),$$

which expresses the desired pd of behavior $b \equiv (s_0, a_1, s_1, a_2, s_2, \ldots, a_T, s_T) \in \mathbb{B}$. It sets high probability values to preferred behaviors and low probability values to unwanted behaviors. It consists of an ideal model m^i of the system and of an ideal decision rule r^i. The real pd $\mathsf{c}^\pi(b)$ depends on the model m of the system and decision rules r forming the policy π.

$$\mathsf{c}^\pi(b) = \prod_{t\in\mathbb{T}} \mathsf{m}(s_t|a_t, s_{t-1})\mathsf{r}(a_t|s_{t-1}).$$

This paper exploits Bayes' learning to get m relating (the observed state, the used action, the next state). The optimal policy π^o in a set $\mathsf{\Pi}$ minimizes the Kullback-Leibler divergence (KLD) of the pd c^π to the ideal pd c^i

$$\pi^o \in \mathsf{Arg}\min_{\pi\in\mathsf{\Pi}} \mathsf{D}(\mathsf{c}^\pi||c^i) = \mathsf{Arg}\min_{\pi\in\mathsf{\Pi}} \int_{b\in\mathbb{B}} \mathsf{c}^\pi(b) \ln\left(\frac{\mathsf{c}^\pi(b)}{\mathsf{c}^i(b)}\right)\mathsf{d}b.$$

Theorem 1. *(FPD, [22]) Decision rules, which constitute the optimal decision policy π^o, are computed for $t = T, T-1, \ldots, 1$ and with $\mathsf{h}(s_T) \equiv 1$ as follows*

$$\mathsf{r}^o(a_t|s_{t-1}) \equiv \mathsf{r}^i(a_t|s_{t-1})\frac{\exp[-\mathsf{d}(a_t, s_{t-1})]}{\mathsf{h}(s_{t-1})},$$

$$\mathsf{d}(a_t, s_{t-1}) \equiv \int_{s_t\in\mathbb{S}} \mathsf{m}(s_t|a_t, s_{t-1}) \ln\left[\frac{\mathsf{m}(s_t|a_t, s_{t-1})}{\mathsf{h}(s_t)\mathsf{m}^i(s_t|a_t, s_{t-1})}\right] \mathrm{d}s_t \tag{1}$$

$$\mathsf{h}(s_{t-1}) \equiv \int_{a_t\in\mathbb{A}} \mathsf{r}^i(a_t|s_{t-1}) \exp[-\mathsf{d}(a_t, s_{t-1})]\, \mathrm{d}a_t \in \mathsf{h}(s_t) \in [0, 1].$$

$$\textit{The attained minimum is } \min_{\pi\in\mathsf{\Pi}} \mathsf{D}(\mathsf{c}^\pi||\mathsf{c}^i) = -\ln(\mathsf{h}(s_0)). \tag{2}$$

We focus on the preference quantification, on finding the c^i. The preference specification is mostly incomplete due to the agent's imperfections. This means that

$$\mathbb{C}^i \equiv \{\text{ideal pds } \mathsf{c}^i(b), b \in \mathbb{B}, \text{respecting the agent's wishes}\} \tag{3}$$

includes several pds. It can be also empty because of the agent's inconsistencies. The agent's preferences can be in contradiction or the agent can have un-achievable goals. The preference elicitation (PE) consists of the choice of: ▶ the non-empty set $\mathbb{C}^i$ that overcomes the agent's inconsistencies ▶ the optimal ideal pd c^{io} from the set (3).

The PE principle from [18] recommends to choose as the optimal ideal pd

$$\mathsf{c}^{io} \in \mathsf{Arg} \min_{\mathsf{c}^i \in \mathbb{C}^i} \min_{\pi \in \Pi} \mathsf{D}(\mathsf{c}^{\pi} || \mathsf{c}^i). \tag{4}$$

Its use in FPD ensures that no preferences are added to the agent's. Theorem 1 describes the 1^{st} minimization over π. The 2^{nd} minimization over c^i is harder and it can be done over individual factors of c^i for each already observed state.

Then, cf. (1), (2), (4), the optimal closed-loop ideal pd c^{io} in the last step reads

$$\mathsf{c}^{io} \equiv \mathsf{m}^{io}\mathsf{r}^{io} \overset{(1),(2)}{\in} \mathsf{Arg} \max_{\mathsf{r}^i \in \mathbb{R}^i} \left[\max_{\mathsf{m}^i \in \mathbb{M}^i} \int_{a_1 \in \mathbb{A}} \mathsf{r}^i(a_1|s_0) \exp[-\mathsf{d}(a_1, s_0)] \mathsf{d}a_1 \right]$$

$$\mathsf{d}(a_1, s_0) = \int_{s_1 \in \mathbb{S}} \mathsf{m}(s_1|a_1, s_0) \ln\left(\frac{\mathsf{m}(s_1|a_1, s_0)}{\mathsf{h}(s_0)\mathsf{m}^i(s_1|a_1, s_0)} \right) \mathsf{d}s_1, \tag{5}$$

$\mathsf{h}(s_0)$ comes from the backward recursion via step (1). The minimization over a c^i- factor $(\mathsf{c}^{\mathsf{i}}(s_t|a_t, s_{t-1}) = \mathsf{m}^{\mathsf{i}}(s_t|a_t, s_{t-1})\mathsf{r}(a_t|s_{t-1}))$ in any decision epoch $t \in \mathbb{T}$ and for any realized state s_{t-1} are formally identical. Therefore, we can suppress t and $s_{t-1} \in \mathbb{S}$ and deal with $\mathsf{m}(s|a) \equiv \mathsf{m}(s_t = s|a_t = a, s_{t-1})$, $\mathsf{m}^i(s|a) \equiv \mathsf{m}^i(s_t = s|a_t = a, s_{t-1})$, $\mathsf{r}(a) \equiv \mathsf{r}(a_t = a|s_{t-1})$, $\mathsf{r}^i(a) \equiv \mathsf{r}^i(a_t = a|s_{t-1})$ and $\mathsf{h}(s) = \mathsf{h}(s_t = s)$. The optimization (5) uses the given $\mathsf{h}(s)$ and runs over $\mathbb{M}^i$ (a set of m^i-s) while $\mathbb{C}^i$ is determined by a given r^i and chosen from the set $\mathbb{R}^i$ (a set of r^i-s). For then $\mathsf{c}^i = \mathsf{m}^i\mathsf{r}^i$- factors are in

$$\{\mathsf{c}^i(s, a) : \mathsf{c}^i(s, a) = \mathsf{m}^i(s|a)\mathsf{r}^i(a), s \in \mathbb{S}, a \in \mathbb{A}, \text{respecting the agent's wishes}\}. \tag{6}$$

2. Preference Quantification

We first perform the optimization for a quite general choice of sets $\mathbb{M}^i, \mathbb{R}^i$. Then, we specialize it to a specific but still general case.

2.1. The generic choice of optimal ideal model of the system

Theorem 2. ***(Optimal** m^{io}**-factor, [19])*** *Let* $\mathsf{r}^i \in \mathbb{R}^i$ *be a fixed ideal decision rule, which defines a non-empty cross-section* $\mathbb{M}^i \equiv \{\mathsf{m}^i : \mathsf{m}^i\mathsf{r}^i \in$ *set* $(6)\}$*. Let* $\mathsf{m}^i(s|a) \in \mathbb{M}^i$ *exist such that* $\mathsf{d}(a) < \infty, \forall a \in \mathbb{A}$ *(1) t and s_{t-1} suppressed. Then, the optimal ideal* m^{io}*−factor minimises* $\mathsf{d}(a), s \in \mathbb{S}, a \in \mathbb{A}$, *i.e.*

$$\mathsf{m}^{io}(s|a) \in \mathsf{Arg} \max_{\mathsf{m}^i \in \mathbb{M}^i} \int_{\mathbb{A}} \mathsf{r}^i(a) \exp[-\mathsf{d}(a)] \mathsf{d}a = \mathsf{Arg} \min_{\mathsf{m}^i \in \mathbb{M}^i} \mathsf{d}(a). \tag{7}$$

2.2. The generic choice of optimal ideal decision rule

The decision rules work on the set of admissible actions. Thus, the support of an admissible r-factor must be included in the set of possible actions i.e. supp[r] $\subset \mathbb{A}$. The form of the FPD-optimal r^o-factor, Theorem 1, implies that supp[r^o] $\subseteq$ supp[r^i]. Therefore, only the ideal r^i-factors

$$\mathsf{r}^i \in \mathbb{R}^i \equiv \left\{\mathsf{r}^i : \text{supp}[\mathsf{r}^i] = \mathbb{A}\right\} \tag{8}$$

keep actions $a \in \mathbb{A}$ and exclude none. Thus, (8) is the generic constraint and

$$\mathbb{R}^i \equiv \{\mathsf{r}^i : \mathsf{m}^{io}\mathsf{r}^i \in (6) \text{ while } \mathsf{m}^{io} \text{ is given by } (7)\}.$$

Theorem 3. *(**Optimal r^{io}-factor meeting (8), [19]**) Let assumptions of Theorem 2 hold and for a scalar $p > 1$*

$$\mathbb{R}^i \equiv \left\{\mathsf{r}^i : supp[\mathsf{r}^i] = \mathbb{A},\ ||\mathsf{r}^i||_p \equiv \left[\int_{\mathbb{A}} (\mathsf{r}^i(a))^p \mathsf{d}a\right]^{1/p} < \infty \right\},\ |\mathbb{A}| \equiv \int_{\mathbb{A}} \mathsf{d}a < \infty. \tag{9}$$

Then, the optimal ideal r^{io}-factor reads, cf. (1), (7),

$$\begin{aligned} \mathsf{r}^{io} &\propto \chi_{\mathbb{A}}(a)\exp[-\nu \mathsf{d}^o(a)], \quad \nu \equiv \frac{1}{p-1},\ \chi_{\mathbb{A}}(a) \textit{ is the indicator function of } \mathbb{A} \\ \mathsf{d}^o(a) &\equiv \int_{\mathbb{S}} \mathsf{m}(s|a)\ln\left(\frac{\mathsf{m}(s|a)}{\mathsf{h}(s)\mathsf{m}^{io}(s|a)}\right)\mathsf{d}s \overset{(7)}{\leq} \mathsf{d}(a). \end{aligned} \tag{10}$$

The r^{io}-factor (10) belongs to (9) and meets (8).

Remarks ▶ The generic constraint (8) implies that the ideal r^i-factors support exploration, which makes the Bayesian learning efficient. ▶ The parameter ν controls exploration. Every action from the set of possible actions can be tried with almost the same probability if the parameter ν is close to 0. If ν gets bigger the exploration declines, cf. form of r^{io} in (10).

2.3. The specific choice of $\mathbb{M}^i$ making $\mathbb{C}^i \neq \emptyset$

The optimal ideal r^{io}-factor is uniquely given by the choice of m^{io} (and by the opted ν) via (10). The description of the agent's preferences only guarantees a non-empty set $\mathbb{M}^i$. A wide range of practical cases can be covered with a few additional PE-oriented queries. Our specific elaborated case concerns the next agent's general wish.

The agent wants to reach given sets of ideal states $\mathbb{S}^i$ and ideal actions $\mathbb{A}^i$,

$$\emptyset \neq \mathbb{S}^i \subset \mathbb{S},\ \emptyset \neq \mathbb{A}^i \subseteq \mathbb{A}. \tag{11}$$

This is quantified as the wish to assign the highest probability to the set of ideal states $\mathbb{S}^i$ and to the set of ideal actions $\mathbb{A}^i$ (11) by closing the loop of the given model m and of the optimal ideal decision rule r^{io}. So we choose as the maximized functional

$$\int_{\mathbb{A}} \rho(a)\mathsf{r}^{io}(a)\,\mathrm{d}a \equiv \int_{\mathbb{A}} \left[(1-w)\int_{\mathbb{S}} \chi_{\mathbb{S}^i}(s)\mathsf{m}(s|a)\,\mathrm{d}s + w\chi_{\mathbb{A}^i}(a)\right]\mathsf{r}^{io}(a)\,\mathrm{d}a. \tag{12}$$

The introduced weight $w \in \mathbb{W} \equiv [0,1]$ parameterizes how much the agent prefers to stay in the set of ideal actions $\mathbb{A}^i$ relative to being in the set of ideal states $\mathbb{S}^i$.

The inspected problem has a meaningful solution if

$$\rho(a) = (1-w)\int_{\mathbb{S}} \chi_{\mathbb{S}^i}(s)\mathsf{m}(s|a)\,\mathrm{d}s + w\chi_{\mathbb{A}^i}(a) > 0, \text{ on } \mathbb{A}. \tag{13}$$

If the functional (12) is large, then the probabilities of the preferred sets are large. The part $(1-w)\int_{\mathbb{S}} \chi_{\mathbb{S}^i}(s)\mathsf{m}(s|a)\,\mathrm{d}s$ forces the highest probability to the set $\mathbb{S}^i$. And the part $w\chi_{\mathbb{A}^i}(a)\mathsf{r}^{io}(a)$ should guarantee that the ideal decision rule will often choose the actions from the set $\mathbb{A}^i$. The weight w balances these probabilities.

Remarks ▶ The weight is fixed. Its fine-tuning is controlled by additional queries. ▶ The function determining $\rho(a)$ qualitatively plays the role of the reward. ▶ Our construction of the optimal ideal pd c^{io} quantifies the agent's preferences in an ambitious but realistic way. ▶ Maximization of (12) with r^{io} given by (10) rely on:

Theorem 4. ***(Optimal value of*** d^o***, [19])*** *Under assumptions of Theorem 3, covering those of Theorem 2, and under (13), the optimal ideal model* m^{io} *fulfilling (12) determines* $\mathsf{d}^o(a)$*, giving* $\mathsf{r}^{io} = \mathsf{r}^i(\mathsf{m}^{io})$ *(10),* $a \in \mathbb{A}$*, as the function*

$$\mathsf{d}^o(a) \quad \equiv \quad \mathsf{d}^o(\bar{a}) + \ln\left(\frac{\max_{a\in\mathbb{A}}(\rho(a))}{\rho(a)}\right), \; \bar{a} \in \mathrm{Arg}\max_{a\in\mathbb{A}}(\rho(a)). \tag{14}$$

Theorem 5. ***(Solvability of (14), [19])*** *Under (13) and* $|\mathbb{A}| < \infty$*, the smallest* $\mathsf{d}^o(\bar{a})$ *exists such that (14) has a solution* $\mathsf{m}^{io}(s|a)$*,* $s \in \mathbb{S}$*,* $\forall a \in \mathbb{A}$*. Thus, the smallest* $\mathsf{d}^o(\bar{a})$ *guaranteeing solvability of (14)* $\forall a \in \{a\}$ *is*

$$\mathsf{d}^o(\bar{a}) = \max\left[0, \max_{a\in\mathbb{A}} \int_{\mathbb{S}} \mathsf{m}(s|a) \ln\left[\frac{\rho(a)}{\rho(\bar{a})\mathsf{h}(s)}\right] \mathrm{d}s\right]. \tag{15}$$

The ideal m^{io} gives $\mathsf{d}^o(a)$ (1) and $\mathsf{r}^{io}(\mathsf{m}^{io})$ via (7). The next proposition provides it for generic pds $\mathsf{m}(s|a)$. It requires to find m^{io} giving d^o (14) on $\mathbb{A}$.

Theorem 6. ***(*** m^{io} ***meeting (12), generic*** $\mathsf{m}(s|a)$***, [19])*** *Let* $\mathsf{m}(s|a)$*, for some* $a \in \mathbb{A}$*, be non-uniform on* $\mathbb{S}$ *and Theorem (3) hold. Then, the* m^{io}*–factor meeting (12) reads*

$$\mathsf{m}^i(s|a) \quad = \quad \frac{\mathsf{m}(s|a)\exp(-\mathsf{e}(a)\mathsf{m}(s|a))}{\int_{\mathbb{S}} \mathsf{m}(s|a)\exp(-\mathsf{e}(a)\mathsf{m}(s|a))\,\mathrm{d}s}, \; \textit{while } |\mathbb{S}| \equiv \int_{\mathbb{S}} \mathrm{d}s < \infty. \tag{16}$$

The real valued $\mathsf{e}(a)$ *in (16) is the existing solution of* $\mathsf{L}(\mathsf{e}(a)) = \mathsf{R}(a)$*. For* $\mathsf{d}^o(\bar{a})$ *meeting (15) with* $\bar{a} \in \mathrm{Arg}\max_{a\in\mathbb{A}} \rho(a)$*, the left- and right-hand sides of this equation are*

$$\mathsf{L}(\mathsf{e}(a)) \quad \equiv \quad \mathsf{e}(a)\Lambda(a) + \ln\left(\int_{\mathbb{S}} \mathsf{m}(s|a)\exp[-\mathsf{e}(a)\mathsf{m}(s|a)]\,\mathrm{d}s\right), \; \Lambda(a) \equiv \int_{\mathbb{S}} \mathsf{m}^2(s|a)\,\mathrm{d}s$$

$$\mathsf{R}(a) \equiv -\int_{\mathbb{S}} \mathsf{m}(s|a) \ln\left(\frac{\mathsf{m}(s|a)}{\mathsf{h}(s)}\right) \mathrm{d}s + \mathsf{d}^o(\bar{a}) + \ln\left(\frac{\rho(\bar{a})}{\rho(a)}\right), \; \bar{a} \in \mathrm{Arg}\max_{a\in\mathbb{A}} \rho(a). \tag{17}$$

The uniform case was solved similarly, see [19].

3. On algorithmization

In the considered case with the discrete-valued states and actions, the found solution can be directly converted into a compact algorithm. It is done in [19]. Here, we just stress that it uses the Bayesian estimation of unknown but time-invariant values of the transition probabilities Θ. The gained parametric model $\mathsf{m}(s_t|a_t, s_{t-1}, \Theta)$ belongs to the exponential family [1] and makes Dirichlet's prior pd self-reproducing. Its degrees of freedom counting the observed transitions $s_{t-1} = \tilde{s} \in \mathbb{S}$, $a_t = a \in \mathbb{A}$ to $s_t = s \in \mathbb{S}$ form the sufficient statistic for learning unknown $\Theta_{s|a,\tilde{s}} \equiv \mathsf{m}(s|a, \tilde{s}, \Theta)$ [3].

4. Dialogue with the user

The agent specifies the preferred states $\mathbb{S}^i$ and preferred actions $\mathbb{A}^i$ before the beginning DM. A problem arises as the agent[1] wishes concern two usually contradiction things. In this case, we need to choose the weight w in (12), which determines how much the user prefers to stay in the set $\mathbb{A}^i$ relative to being in $\mathbb{S}^i$. But they are unable to express how much they prefer it before they will observe how the closed loop behaves. That is why we added a dialogue with the user during the DM. The user will express their preferences and next they will control the results of the DM during the DM. The DM solved in Section 3, referred to as the basic DM, deals with two types of inputs:

- ✓ those directly describing the basic DM, which include: ▶ the state $\mathbb{S}$ and action $\mathbb{A}$ sets; ▶ the wishes-expressing ideal sets $\mathbb{S}^i \subset \mathbb{S}$ and $\mathbb{A}^i \subseteq \mathbb{A}$;
- ✓ more technical, policy-influencing, inputs that include: ▶ the weight $w \in [0, 1]$ balancing the relative importance of ideal sets, see (12); ▶ the scalar $\nu > 0$, see (10), balancing exploration with exploitation (duality, [10, 20]).

Fine variations of ideal sets $\mathbb{S}^i$, $\mathbb{A}^i$ or the design horizon $|T|$ are potential inputs of the preference processing but they are here fixed. Thus, the paper focuses just on the pair w, ν. Its optimal choice depends on: ▶ subjective user's preferences; ▶ the user's attitude to the basic DM; ▶ emotions, etc., i.e. on the user's mental state. The dependence is complex and the mental state can hardly be directly measured and quantified. Two users can have the same preferences expressed by the sets $\mathbb{S}^i$, $\mathbb{A}^i$, but their responses differ.

In our solution, the user is asked to judge the DM quality reached for various choices of w, ν. This is the domain of classical PE [8] that often elicits preferences about a static DM and interactively queries the user. Even advanced versions, represented by [4, 5, 7], become cumbersome in the targeted basic *dynamic* DM. This makes us adopt the next user-driven way that consists of solving an appropriate FPD meta-task, whose description uses capital versions of all functions and parameters entering it, cf. [11].

The user assigns (satisfaction) marks. Their changes during the dialog serve as the (meta-)state $S_T \in \bar{\mathbb{S}}$, to the behaviour caused by the policy, designed for trial values of the optional inputs here, (w, ν). Their changes A_T are the (meta-)actions. They are generated by (meta-)policy gained by the same algorithm as that used at the basic level[2]. It runs more slowly than the basic DM, $T \in \{\bar{T}, 2\bar{T}, \ldots, \} \subset \mathbb{T}$ given by $\bar{T} > 1$.

This simple idea has to cope with the possible infinite regress, i.e. DM at meta-level needs meta-inputs opted via a meta-PE, etc. Also, the curse of dimensionality [2] endangers applicability as the opted inputs are multiple and continuous-valued. Our way counteracts both obstacles. We

[1] The agent will be called user as it is usual for preference elicitation.

[2] In harmony with the quest for a universal DM.

decided to ask queries after every time epoch $\bar{T} > 1$, but the queries can be answered irregularly after some multiples of the $\bar{T}$. The use of zero-order holder copes with the expected irregularity of user's responses. It makes realistic the time-invariance of the model $\mathsf{M}(S_T|A_T, S_{T-\bar{T}}, \Theta) := \Theta_{S_T|A_T,S_{T-\bar{T}}}$ needed for learning this meta-model, cf. the beginning of Sec. 3.

The set of possible meta-states is $\bar{\mathbb{S}} := \{-1, 0, 1\}$. It is implied by a difference of the current mark and previous mark[3] i.e. $\Delta g = g_T - g_{T-1}$. If $\Delta g < 0 \Longrightarrow \bar{\mathbb{S}} = \{-1\}$, if $\Delta g = 0 \Longrightarrow \bar{\mathbb{S}} = \{0\}$ and if $\Delta g > 0 \Longrightarrow \bar{\mathbb{S}} = \{1\}$.

The choice of the ordinal scale of marks $g \in \bar{\mathbb{G}} \equiv \{1, \ldots, |\bar{\mathbb{G}}| \equiv 5\}$ suffices for expressing "satisfaction degree". A rich, cross-domain, experience, e.g. in marketing [6] or in European Credit and Accumulation System, confirms this. The mark $g = 1$ is taken as the best one. The ideal set of meta-states is then $\bar{\mathbb{S}}^i \equiv \{-1\}$.

By construction, the outcomes of the basic DM depend smoothly on the discussed inputs. Thus, changes $A \equiv (\Delta w, \Delta \nu)$ of inputs (w, ν) can be selected in a finite set $\bar{\mathbb{A}} := \{(\Delta w, \Delta \nu)\}$ of discrete values. The natural flexible options are

$$\Delta w \in \{-\bar{w}, 0, \bar{w}\},\ \Delta \nu \in \{-\bar{\nu}, 0, \bar{\nu}\}, \quad \bar{w}, \bar{\nu} > 0. \tag{18}$$

The meta-policy is to guarantee that its actions stay within their allowed ranges ($w \in [0, 1]$, $\nu > 0$). The used simple clipping at boundaries of (18) seems to suffice. We have no other demands on the actions. Thus, $\bar{\mathbb{A}} = \bar{\mathbb{A}}^i$ and $W = 0$ (meta-twin to w in (12)).

The last input to the meta-DM is the parameter of exploration ν. It makes no sense to choose a different value at the meta-level: the meta-action is its common value.

The appearance of $\bar{T}$, $\bar{w}$, $\bar{\nu}$ still preserves the danger of infinite regress. At present, it is cut by force and they are chosen heuristically. They, however, offer, the first step in a conceptual solution that: ▶ lets appear only meta-inputs that have a weak influence on results; ▶ tunes them via an adaptive minimization of miss-modelling error [17].

5. Experiments

This core section presents experiments. We have chosen a DM example with a heating system.

Common simulation options The simulated system is Markov with $|\mathbb{S}| = 15$ and $|\mathbb{A}| = 7$. It is created by learning the transition pd $\mathsf{p}(s_t|a_t, s_{t-1})$ on the simulated system generating 10^6 real values y_t stimulated by independently generated discrete actions in $\mathbb{A} := \{1, \ldots, 7\}$. The states $s_t \in \mathbb{S} := \{1, \ldots, 15\}$ are gained via an affine mapping of discretized values of the real-valued y_t generated by the equation ($y_0 = 1$)

$$y_t = 0.028 y_{t-1} + 1.81 y_{t-2} - 0.817 y_{t-3} + 0.1 a_t - 0.16 a_{t-1} + 0.05 \varepsilon_t.$$

There, ε_t is the white, zero-mean, normal noise with a unit variance. In all experiments with the Markov chain, the number of simulated epoch was 800. The seed of the random generator was fixed, and the initial state $s_0 = 1$. The initial guess of the entries of the array e (17) was 1.2. The horizon for dynamic programming is $h = 2$, which suffices when taking the outcome from the previous epoch as the initial guess of the stationary value function.

[3] We decided to note marks with a symbol g as grade, as m for mark is already used.

Experiments We present DM results without and with the user's control. DM without the user's control, it is the basic DM with no meta-level and preferences expressed by the ideal sets $\mathbb{S}^i$, $\mathbb{A}^i$ and by fixed options w, ν. DM with the user's control solves the basic DM supported by the second-layer implementing the solution of the meta-DM task with the dialogue with the user. The DM with the user's control gives the user the chance to express their satisfaction every ten steps, $\bar{T} = 10$. The satisfaction is quite subjective. It is demonstrated by presenting selected results for different users. We also present results with different fixed parameters w, ν to show how these parameters influence DM. In experiments with the user's control, these parameters are free and they are changed by the responses of the user. The changes of the free parameters w, ν are $\bar{w} = 0.1$ and $\bar{\nu} = 0.3$ (18). To compare the results impartially we use prices paid for deviation from the preferred behavior. The agreed prices are in Table 1 (common to all experiments) and Table 2 that suits to the preferred state $\mathbb{S}^i = \{8\}$. Other preferred states are priced similarly.

Table 1

The price paid for the individual action values

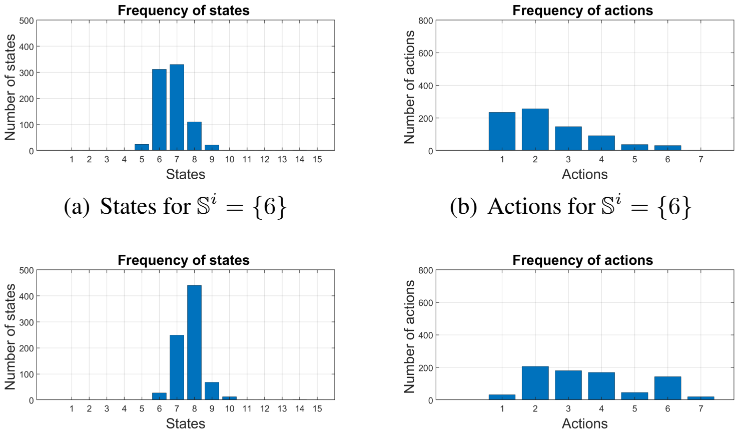
action	1	2	3	4	5	6	7
price	3	2	1	0	1	2	3

Table 2

The price paid for the individual state values when $\mathbb{S}^i = \{8\}$

state	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
price	4	3	3	2	2	1	1	0	1	1	2	2	3	3	4

Experiment 1. It shows the results for the preferred state $\mathbb{S}^i = \{6\}$ and then for $\mathbb{S}^i = \{8\}$. No action is preferred, $\mathbb{A}^i = \mathbb{A}$. The free parameters are fixed, $w = 0, \nu = 1$.



(a) States for $\mathbb{S}^i = \{6\}$ (b) Actions for $\mathbb{S}^i = \{6\}$

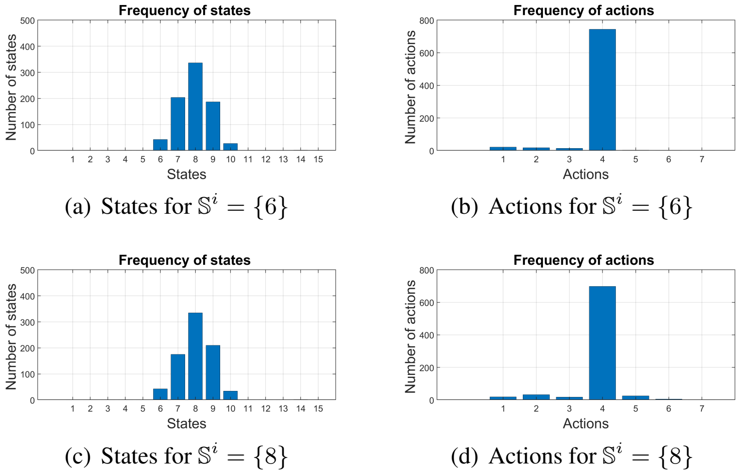
(c) States for $\mathbb{S}^i = \{8\}$ (d) Actions for $\mathbb{S}^i = \{8\}$

Figure 1: Exp. 1: states and actions in the basic DM for different preferences of states without any additional preference of actions $\mathbb{A}^i = \mathbb{A}$

Discussion In the Fig. 1, we can see that the frequency of the preferred state $\mathbb{S}^i = \{8\}$ is pretty high. It occurs the most often. On the other hand the preferred state $\mathbb{S}^i = \{6\}$ does not occur the most often and its frequency is low. It is hard for the system to get the state $\mathbb{S}^i = \{6\}$. We will try to change the free parameters to improve results.

Experiment 2. It shows the results for $\mathbb{S}^i = \{8\}$ and $\mathbb{S}^i = \{6\}$ with the extra preference of actions $\mathbb{A}^i = \{4\}$. The weight $w = 0.3$ and the value $\nu = 1$ are fixed.

Discussion With the extra preference on actions the preferred states appear less often for both preferences, but still the preferred state appears the most often for the preference $\mathbb{S}^i = \{8\}$. On



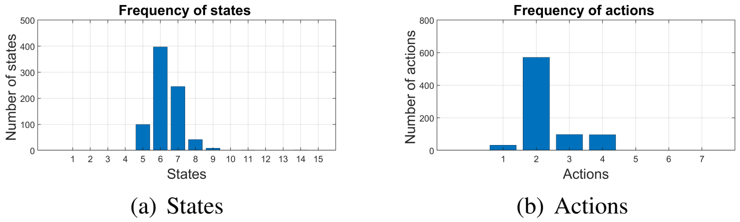
(a) States for $\mathbb{S}^i = \{6\}$ (b) Actions for $\mathbb{S}^i = \{6\}$

(c) States for $\mathbb{S}^i = \{8\}$ (d) Actions for $\mathbb{S}^i = \{8\}$

Figure 2: Exp. 2: states and actions in the basic DM for different preferences of states with an extra preference on actions $\mathbb{A}^i = \{4\}, w = 0.3$

the other hand the results for actions are pretty good for both. The preferred state $\mathbb{S}^i = \{6\}$ with the extra preference of action occurs even much less often, as expected, because these preferences contradict. But still the results are not bad. The user can be satisfied with the results because they could prioritized the results of actions over the poorer results concerning states.

Experiment 3. We would like to show how the parameter ν influence the results. We try improve our results for the preferred state $\mathbb{S}^i = \{6\}$, which gives worse results. We should be able to improve the results when there is no additional preference of actions. If the exploration parameter ν will be bigger, the selection of the action will not be uniform, but will be concentrated on the action, which guarantees the preferred state.

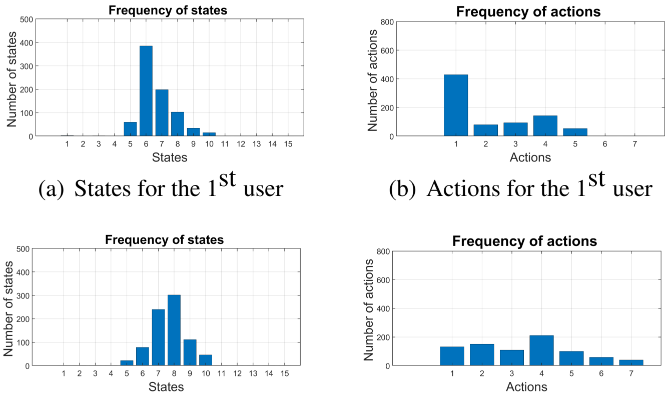


(a) States (b) Actions

Figure 3: Exp. 3: states and actions in the basic DM for $\mathbb{S}^i = \{6\}, \mathbb{A}^i = \mathbb{A}$, $\nu = 5$

Discussion We can see, that we can improve the previous results via the parameter ν. We tried many values of ν, and present the best of for which the preferred state occurs the most often. The actions that cause the state 6 are around the action 2.

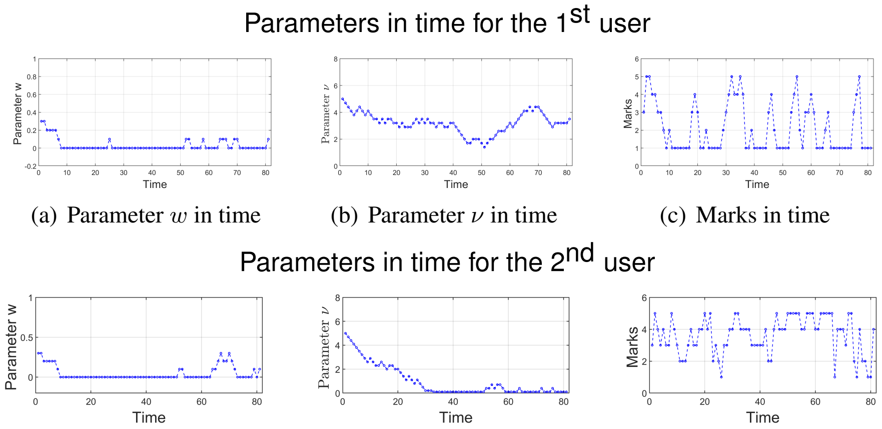
Experiment 4. We showed that the results of the experiments are influenced by the parameters w and ν. That is why we left these parameters to be free for the dialogue with the user. We choose their values according to the responses of the users. We would like to show, that they can get the desired results without any knowledge of DM and PE theories just using our algorithm. The users were instructed to want $\mathbb{S}^i = \{6\}$ without an additional preference of actions $\mathbb{A}^i = \mathbb{A}$ and then $\mathbb{S}^i = \{8\}$ with their preference of actions $\mathbb{A}^i = \{4\}$.



(a) States for the 1$^{\text{st}}$ user (b) Actions for the 1$^{\text{st}}$ user

(c) States for the 2$^{\text{nd}}$ user (d) Actions for the 2$^{\text{nd}}$ user

Figure 4: Exp. 4: states and actions for $\mathbb{S}^i = \{6\}, \mathbb{A}^i = \mathbb{A}$ in DM for the users



Parameters in time for the 1$^{\text{st}}$ user

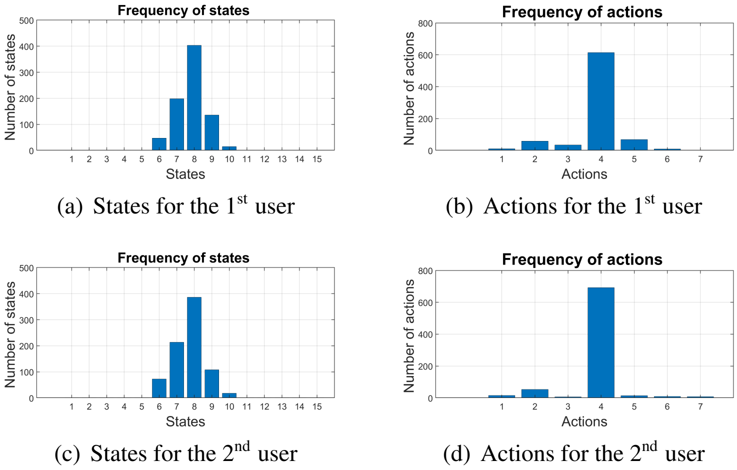
(a) Parameter w in time (b) Parameter ν in time (c) Marks in time

Parameters in time for the 2$^{\text{nd}}$ user

(d) Parameter w in time (e) Parameter ν in time (f) Marks in time

Figure 5: Exp. 4: The evolution of parameters for users with preferences $\mathbb{S}^i = \{6\}, \mathbb{A}^i = \mathbb{A}$.

Discussion In the Fig. 4 we can see that the results for $\mathbb{S}^i = \{6\}, \mathbb{A}^i = \mathbb{A}$ are much better for the 1$^{\text{st}}$ user. The occurrence of the preferred state is pretty high and it appears the most often. For the 2$^{\text{nd}}$ user the results are worse. The preferred state does not appear as much often and its occurrence is low. If we look at the evolution of the free parameters and marks Fig. 5, we can see that the courses of weight w are very similar and the weight is zero most of the time, which we assumed, because there is no additional preference. The courses of parameter ν are pretty different. The 1$^{\text{st}}$ user's ν is almost all the time between 2 and 5 and the 2$^{\text{nd}}$ user's declines to the value 0.1. That is because the 2$^{\text{nd}}$ user was strict with their marking, they were not so satisfied and that's why the algorithm tries to increase the exploration and find the results to satisfy the user. Because of that the results got worse and we got into dead end. So it really depends on the user's strategy. Thanks to the 1$^{\text{st}}$ user we can see that we can get good results but also thanks to the 2$^{\text{nd}}$ user we can see that if the algorithm gets bad feedback, it will worsen the results.



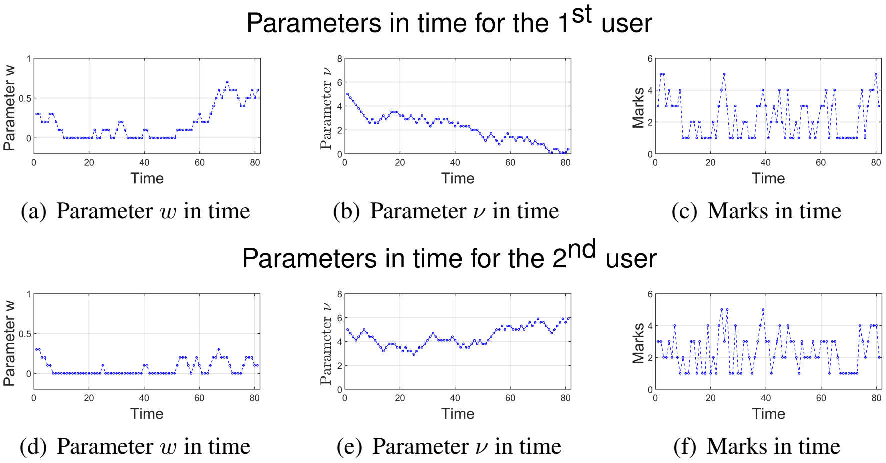
(a) States for the 1st user

(b) Actions for the 1st user

(c) States for the 2nd user

(d) Actions for the 2nd user

Figure 6: Exp. 4: states and actions for $\mathbb{S}^i = \{8\}, \mathbb{A}^i = \{4\}$ in DM for the user



(a) Parameter w in time

(b) Parameter ν in time

(c) Marks in time

(d) Parameter w in time

(e) Parameter ν in time

(f) Marks in time

Figure 7: Exp. 5: The evolution of parameters with preferences $\mathbb{S}^i = \{8\}, \mathbb{A}^i = \{4\}$.

Discussion For the preferences $\mathbb{S}^i = \{8\}, \mathbb{A}^i = \{4\}$, Fig. 5 the both users got great results. These preferences are not in a contradiction and as we could see above, it is easy for the system to reach this state. We can see from marking that both users were satisfied. The courses of the weight and ν differ. The 1st user's weight increase more and parameter ν decrease more than for the 2nd user. The 2nd user's parameters are more consistent but the frequencies of preferred state and action do not differ much.

6. Numerical results

Table 3 shows the prices paid for actions and states for $\mathbb{S}^i = \{8\}, \mathbb{A}^i = \{4\}$. For $\mathbb{S}^i = \{6\}, \mathbb{A}^i = \mathbb{A}$ the results can be judged in the same way.

We can see that the total price is the best (the lowest) for the 2nd user because they had the

Table 3
The price paid for actions and states in all experiments

Exp. no	Opted parameters	The price of actions	The price of states	Total price	Number of the preferred actions	Number of the preferred states
1.	$w = 0.0,\ \nu = 1$	1086	370	1456	170	440
2.	$w = 0.3,\ \nu = 1$	181	475	656	698	335
3.	1st user	281	403	684	614	403
4.	2nd user	219	420	639	692	386

top number of selections of the preferred action and the preferred state occurs also very often. They were satisfied as can be seen on the evolution of marks Fig 7. So we can say that it is the best result of our experiments. But the user are different so for someone it is a good results and for someone else not, because the prices that the users are willing to pay are individual. That we should keep in mind. The ”objective” numerical comparison is of secondary importance. We also repeat that the users got the results they want without any knowledge of DM theory. It is also less time demanding to find a good policy via their feedback during the DM.

7. Concluding remarks

The paper presents the quantification of preferences within the fully probabilistic design of decision results. It provides the user's feedback that optimizes free parameters ν and w. It presents the experiments which show how the fixed parameters influence the DM. It compares the DM with and without the user's control. The algorithm does not need users any additional knowledge of the DM and PE theories.

The further research should:

- ✓ care about dimensionality curse connected with other wishes;
- ✓ add more free parameters, e.g., extensions of preferred sets of states and actions;
- ✓ address continuous systems;
- ✓ more specific application and real-system cases; etc.

These are hard tasks requiring more research to fill the gaps in the built universal DM theory, cf. **Motivation**.

Acknowledgement: *We would like to thank for the support of this paper by grant MŠMT SGS21/165/OHK4/3T/14 and by the department of Adaptive Systems of Institute of Information Theory and Automation of the Czech Academy of Sciences.*

References

[1] Barndorff-Nielsen, O.: Information and Exponential Families in Statistical Theory. Wiley, N.Y. (1978)
[2] Bellman, R.: Adaptive Control Processes. Princeton U. Press, NJ (1961)
[3] Berger, J.: Statistical Decision Theory and Bayesian Analysis. Springer (1985)

[4] Boutilier, C.: A POMDP formulation of preference elicitation problems. In: Proc. of the 18th National Conf. on AI, AAAI-2002. pp. 239–246. Edmonton, AB (2002)

[5] Boutilier, C., et al: A constraint-based approach to preference elicitation and decision making. In: Proc. of AAAI Spring Symp. on Qualitative Decision Theory, pp. 19–28. AAAI Press (1997)

[6] Brace, I.: Questionnaire design. How to plan, structure and write survey material for effective market research. Kogan Page, London (2004)

[7] Braziunas, D., Boutilier, C.: Elicitation of factored utilities. AI Magazine **29**(4), 79–91 (2008)

[8] Chen, L., Pu, P.: Survey of preference elicitation methods. Tech. Rep. IC/2004/67, HCI Group Ecole Politechnique Federale de Lausanne, Switzerland (2004)

[9] Dwivedi, Y., et al: Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy

[10] Feldbaum, A.: Theory of dual control. Autom. Remote Control **22**, 3–19 (1961)

[11] Greco, C., Suglia, A., Basile, P., Semeraro, G.: Converse-et-impera: Exploiting deep learning and hierarchical reinforcement learning for conversational recommender systems. In: Conf. of the Italian association for artificial intelligence. pp. 372–386. Springer (2017)

[12] Guy, T., Kárný, M., Rios-Insua, D., Wolpert, D. (eds.): Proc. of the NIPS 2016 Workshop on Imperfect Decision Makers, vol. 58. JMLR (2016)

[13] Guy, T., Kárný, M., Wolpert, D.: Decision Making and Imperfection, vol. 474. Springer, Berlin (2013)

[14] Guy, T., Kárný, M., Wolpert, D.: Decision Making: Uncertainty, Imperfection, Deliberation and Scalability, vol. 538. Springer, Switzerland (2015)

[15] Hutter, M.: Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability. Springer, Berlin, Heidelberg, N.Y. (2005)

[16] Kárný, M.: Fully probabilistic design unifies and supports dynamic decision making under uncertainty. Inf. Sci. **509**, 104–118 (2020)

[17] Kárný, M.: Towards on-line tuning of adaptive-agent's multivariate meta-parameter. Int. J. of Machine Learning and Cybernetics **12**(9), 2717–2731 (2021)

[18] Kárný, M., Guy, T.: Preference elicitation within framework of fully probabilistic design of decision strategies. In: IFAC Int. Workshop on Adaptive and Learning Control Systems. vol. 52, pp. 239–244 (2019)

[19] Kárný, M., Siváková, T.: Agent's feedback in preference elicitation. In: 2021 20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS). pp. 421–429 (2021). https://doi.org/10.1109/IUCC-CIT-DSCI-SmartCNS55181.2021.00073

[20] Mesbah, A.: Stochastic model predictive control with active uncertainty learning: A survey on dual control. Annual Reviews in Control **45**, 107 – 117 (2018)

[21] Müller, V., Bostrom, N.: Future Progress in Artificial Intelligence: A Survey of Expert Opinion, pp. 555–572. Springer Int. Pub. (2016)

[22] Šindelář, J., Vajda, I., Kárný, M.: Stochastic control optimal in the Kullback sense. Kybernetika **44**(1), 53–60 (2008)