Agent's Feedback in Preference Elicitation*

1st Miroslav Kárný
The Czech Academy of Sciences,
Institute of Information Theory and Automation,
182 00 Prague 8, Czech Republic
Prague, Czech Republic
school@utia.cas.cz,0000-0002-7440-6041

2nd Tereza Siváková
The Czech Academy of Sciences,
Institute of Information Theory and Automation,
182 00 Prague 8, Czech Republic
Prague, Czech Republic
terka.sivakova@seznam.cz

***Abstract*—A generic *d*ecision-*m*aking (DM) agent specifies its preferences partially. The studied prescriptive DM theory, called *f*ully *p*robabilistic *d*esign (FPD) of decision strategies, has recently addressed this obstacle in a new way. The found preference completion and quantification exploits that: ▶ FPD models the closed DM loop and the agent's preferences by joint *p*robability *d*ensities (pds); ▶ there is a *p*reference-*e*licitation (PE) principle, which maps the agent's model of the state transitions and its incompletely expressed wishes on an *ideal* pd quantifying them. The gained algorithmic quantification provides ambitious but potentially reachable DM aims. It suppresses demands on the agent selecting the preference-expressing inputs. The remaining PE options are: ▶ a parameter balancing exploration with exploitation; ▶ a fine specification of the ideal (desired) sets of states and actions; ▶ relative importance of these ideal sets. The current paper makes decisive steps towards a systematic and realistic choice of such inputs by solving a meta-DM task. The algorithmic "meta-agent" observes the user's satisfaction, expressed by school-type marks, and tunes the free PE inputs to improve these marks. The solution requires a suitable formalisation of such a meta-task. This is done here. The proposed way copes with the danger of infinite regress and the dimensionality curse. Non-trivial simulations illustrate the results.**

***Index Terms*—Preference elicitation, Adaptive, agent, Decision making, Bayes' rule**

I. Introduction

An agent opting and using actions to meet its wishes[1] solves *d*ecision *m*aking (DM) task. The choice of an optimal, action-opting, strategy needs the quantification tailored to the used DM theory. The adopted Bayesian paradigm [40] elicits prior beliefs about relations in the closed-loop, formed by the agent and its environment [13], [37], and updates them by (extended) Bayes' rule [4], [30], [38]. The minimum relative-entropy principle [41] completes the probabilistic models. The quantification of the agent's wishes is a harder problem as included humans: ▶ poorly cope with multi-attribute DM tasks [15]; ▶ are prone to contradictions [18]; ▶ spare the deliberation effort on this DM subtask [20].

The paper continues in complementing the still-insufficient support of the *p*reference *e*licitation (PE). It deals with the preference quantification for *dynamic* DM in the vein of [27]–[29]. Similarly, as these works, it processes the state-transition model and a semi-verbal expression of the agent's wishes. The processing delimits the ideal *p*robability *d*ensity (pd) quantifying the agent's wishes. It may initiate the usual PE [12] and simplify the query-based PE [9], [14] as it reduces the amount of tuned PE inputs.

The current paper additionally offers the user the opportunity to express its satisfaction. This serves for adapting the remaining wishes-quantifying inputs. The active querying and processing of the agent's answers [7] is here left aside.

The addressed PE serves to *f*ully *p*robabilistic *d*esign (FPD) of decision strategies [26]. FPD models preferences by the ideal pd describing the *desired* pd of all thought variables (called behaviour). The FPD-optimal strategy makes the behaviour-modelling pd the closest one to its ideal twin. *K*ullback-*L*eibler *d*ivergence (KLD) [32] expresses their closeness. Note that FPD has KL control [22], [44] as its particular case. FPD also densely extends Bayesian DM [23] represented by Markov decision process [11], [16], [19]. PE within FPD consists of: ▶ a translation of the agent's wishes into a non-empty set of imminent ideal pds; ▶ a choice of the optimal ideal pd that adds as little extra wishes or constraints as possible; and ▶ adaptation of inputs entering the previous steps.

The last of the above PE steps is the *main paper topic*.

Layout: Sec. II recalls FPD, the used PE principle and its most advanced elaboration. Core Sec. III describes meta-FPD with this type PE applied to the tuning of free inputs quantifying wishes. Sec. IV illustrates the theory by experiments. Sec. V summarises the results and outlines open issues. The related works are sampled throughout the text.

Notation: $\{x\}$ marks the set of xs. It is a part of a real vector space or of a set of pds. It is detailed if needed. $:=$ defines by assigning. $^{\mathsf{o}}$ marks optimum. $^{\mathsf{i}}$ points to the ideal pd or set. $\mathsf{sansmath}$ fonts mark mappings. $\propto$ is proportionality. $||\mathsf{f}||_p := \left[\int_{\{x\}} |\mathsf{f}(x)|^p \,\mathrm{d}x\right]^{\frac{1}{p}}$, $p > 1$, is L^p norm [39] of a real-valued function $\mathsf{f}(x)$ on $\{x\}$. Integral notation also applies to discrete x. $|x| := \int_{\{x\}} \mathrm{d}x$ is the volume (cardinality) of $\{x\}$. $\chi_{\{x\}}(x)$ is the indicator function of the set $\{x\}$ at x. $\operatorname{Arg}\min_{x\in\{x\}} \mathsf{f}(x) \subset \{x\}$ contains minimisers of $\mathsf{f}(x)$ (the existence is assumed). A known initial state s_0 is implicitly in all conditions. Small letters concern the agent's options. Their capital twins concern

Supported by MŠMT LTC18075 and EU-COST Action CA16228

[1]It means human, device or their group. "Preference" and "wish" serve us as synonyms.

the meta-DM task.

II. Preliminaries

The material presented here should make the paper readable without consulting papers [27]–[29] containing the used and enriched theory.

A. DM via FPD

DM joins the agent and its environment into the closed-loop. The agent uses actions $a_t \in \{a\} \neq \emptyset$ at time $t \in \{t\} := \{1, \ldots, |t|\}$, $|t| \leq \infty$. They influence transitions of the (closed-loop) states $s_{t-1} \in \{s\} \neq \emptyset$ to states $s_t \in \{s\}$. The states and actions up to $|t|$ form the (closed-loop) behaviour $b \in \{b\}$. The agent selects actions via a randomised strategy $r \in \{r\} := \{(r(a_t|s_{t-1}),\, a_t \in \{a\}, s_{t-1} \in \{s\})_{t\in\{t\}}\}$. The pds $r(a_t|s_{t-1})$ (r-factors) are the decision rules forming the strategy r. The r-dependent (closed-loop) model is the joint pd $c^r(b)$ of behaviours $b \in \{b\}$. The chain rule for pds [35] and the state meaning imply

$$c^r(b) = \prod_{t\in\{t\}} m(s_t|a_t, s_{t-1}) r(a_t|s_{t-1}), \tag{1}$$
$$b \in \{b\} := \{b = (s_t, a_t)_{t\in\{t\}}\}.$$

The model $m \in \{m\} := \{(m(s_t|a_t, s_{t-1}),\, s_t, s_{t-1} \in \{s\}, a_t \in \{a\})_{t\in\{t\}}\}$, consists of conditional pds $m(s_t|a_t,\, s_{t-1})$ (m-factors) describing the state transitions.

The m-factors are known. Modelling [6] with Bayesian learning [35] provide them. The state thus includes the used statistic values [17].

FPD quantifies the agent's wishes by an ideal (closed-loop) model. It is a joint pd $c^{\mathrm{i}}(b)$, $b \in \{b\}$, which has high values on preferred behaviours, small on undesired ones and zero on forbidden behaviours. It factorises as the pd (1)

$$c^{\mathrm{i}}(b) = \prod_{t\in\{t\}} m^{\mathrm{i}}(s_t|a_t, s_{t-1}) r^{\mathrm{i}}(a_t|s_{t-1}), \quad b \in \{b\}. \tag{2}$$

The m^{i}- and r^{i}-factors model the desired state transitions and ways of the action choices. The FPD-optimal strategy $r^o \in \{r\}$ minimises KLD $D(c^r||c^{\mathrm{i}})$ of c^r to c^{i}

$$r^o \in \operatorname*{Arg\,min}_{r\in\{r\}} D(c^r||c^{\mathrm{i}})$$
$$:= \operatorname*{Arg\,min}_{r\in\{r\}} \int_{\{b\}} c^r(b) \ln\left(\frac{c^r(b)}{c^{\mathrm{i}}(b)}\right) \mathrm{d}b. \tag{3}$$

The next proposition provides the FPD-optimal strategy (3). Its general case is in [26].

Proposition 1 (FPD): The backward, $t = |t|, |t|-1, \ldots, 1$, functional recursion on $h(s_t) \in [0, 1]$ with $h(s_{|t|}) := 1$ and $s_t \in \{s\}$, $a_t \in \{a\}$,

$$h(s_{t-1}) := \int_{\{a\}} r^{\mathrm{i}}(a_t|s_{t-1}) \exp[-d(a_t|s_{t-1})]\, \mathrm{d}a_t \tag{4}$$
$$d(a_t|s_{t-1}) := \int_{\{s\}} m(s_t|a_t, s_{t-1}) \ln\left[\frac{m(s_t|a_t, s_{t-1})}{h(s_t) m^{\mathrm{i}}(s_t|a_t, s_{t-1})}\right] \mathrm{d}s_t$$

gives the optimal r^o-factors and the value functions $-\ln(h(s_{t-1}))$, [5]. It holds

$$r^o(a_t|s_{t-1}) = \frac{r^{\mathrm{i}}(a_t|s_{t-1}) \exp[-d(a_t|s_{t-1})]}{h(s_{t-1})}, \tag{5}$$
$$\min_{r\in\{r\}} D(c^r||c^{\mathrm{i}}) = -\ln(h(s_0)).$$

B. Optimal PE Principle and Its Use

The ideal pd c^{i} (2) quantifies the agent's wishes. Thus, PE consists of the choice of the pd $c^{\mathrm{i}o}$ that expresses them in the best way. The, generically incomplete, description of wishes delimits the set $\{c^{\mathrm{i}}\}$

$$\{c^{\mathrm{i}}\} := \{\text{ideal pds } c^{\mathrm{i}}(b),\, b \in \{b\}, \tag{6}$$
$$\text{meeting the agent's preferences}\}.$$

The set (6) may be empty due to the agent's inconsistencies or may contain many pds. It may also depend on optional inputs. Thus, PE consists of an amenable choice of:

▶ the *non-empty* set $\{c^{\mathrm{i}}\}$ (6) that copes with inconsistencies of the agent's wishes;
▶ the *optimal* ideal pd $c^{\mathrm{i}o}$ from this set;
▶ the *optional inputs*.

The last choice is the main topic of this paper treated in Sec. III. Here, we recall results solving the initial pair of steps. For $\{c^{\mathrm{i}}\} \neq \emptyset$, which is guaranteed below, the PE principle [27] recommends the choice

$$c^{\mathrm{i}o} \in \operatorname*{Arg\,min}_{c^{\mathrm{i}}\in\{c^{\mathrm{i}}\},\text{see (6)}} \left[\min_{r\in\{r\}} D(c^r||c^{\mathrm{i}})\right]. \tag{7}$$

Obviously, it adds no extra wishes or constraints to those expressed by the agent.

The minimisations over c^{i}-factors ($c^{\mathrm{i}}(s_t, a_t|s_{t-1}) = m^{\mathrm{i}}(s_t|a_t, s_{t-1}) r^{\mathrm{i}}(a_t|s_{t-1})$) at any time $t \in \{t\}$ and for any state s_{t-1} are formally identical. Thus, the description of PE can hide t, s_{t-1} and deal with $m(s|a) := m(s_t = s|a_t = a, s_{t-1})$, $m^{\mathrm{i}}(s|a) := m^{\mathrm{i}}(s_t = s|a_t = a, s_{t-1})$, $r(a) := r(a_t = a|s_{t-1})$, $r^{\mathrm{i}}(a) := r^{\mathrm{i}}(a_t = a|s_{t-1})$ and $h(s) := h(s_t = s)$, $s_{t-1}, s_t, s \in \{s\}$, $a_t, a \in \{a\}$. The optimal $c^{\mathrm{i}o}$-factor, see (4), (5), (7), is then

$$c^{\mathrm{i}o} \in \operatorname*{Arg\,max}_{r^{\mathrm{i}}\in\{r^{\mathrm{i}}\}} \left[\max_{m^{\mathrm{i}}\in\{m^{\mathrm{i}}\}} \int_{\{a\}} r^{\mathrm{i}}(a) \exp[-d(a)]\, \mathrm{d}a\right]$$
$$d(a) = \int_{\{s\}} m(s|a) \ln\left(\frac{m(s|a)}{h(s) m^{\mathrm{i}}(s|a)}\right) \mathrm{d}s,$$
$$d(a) \in \left[-\int_{\{s\}} m(s|a) \ln[h(s)]\, \mathrm{d}s, \infty\right] \tag{8}$$

with $h : \{s\} \to [0, 1]$ gained by the previous design step in (4). The evaluation (8) runs over a cross-section $\{m^{\mathrm{i}}\text{-factors}\}$ of $\{c^{\mathrm{i}}\text{-factors}\}$ given by an r^{i}-factor. Then, it runs over $\{r^{\mathrm{i}}\text{-factors}\}$ for which $c^{\mathrm{i}} = m^{\mathrm{i}} r^{\mathrm{i}}$-factor is a pd on $\{s\}$ and $\{a\}$ in

$$\{c^{\mathrm{i}}\text{-factors}\} := \{m^{\mathrm{i}} r^{\mathrm{i}} = c^{\mathrm{i}}\text{-factor} \tag{9}$$
$$\text{meeting the agent's preferences}\}.$$

The next proposition, proved in [29], provides the optimal ideal m^{io}.

Proposition 2 (Optimal m^{io}-Factor): Let an $\mathsf{r}^{\mathrm{i}} \in \{\mathsf{r}^{\mathrm{i}}\}$ define a non-empty cross-section $\{\mathsf{m}^{\mathrm{i}}\}$ of (9). Let $\mathsf{m}^{\mathrm{i}}(s|a) \in \{\mathsf{m}^{\mathrm{i}}\}$ exist such that $\mathsf{d}(a) < \infty$, $\forall a \in \{a\}$. Then, the optimal ideal m^{io}-factor (8) minimises $\mathsf{d}(a)$, $s \in \{s\}$, $a \in \{a\}$,

$$\mathsf{m}^{\mathrm{io}}(s|a) \in \operatorname*{Arg\,max}_{\mathsf{m}^{\mathrm{i}} \in \{\mathsf{m}^{\mathrm{i}}\}} \int_{\{a\}} \mathsf{r}^{\mathrm{i}}(a) \exp[-\mathsf{d}(a)]\,\mathrm{d}a = \operatorname*{Arg\,min}_{\mathsf{m}^{\mathrm{i}} \in \{\mathsf{m}^{\mathrm{i}}\}} \mathsf{d}(a). \tag{10}$$

The next choice treats universally desirable r^{i}-factors.

The support $\mathrm{supp}[\mathsf{r}^{\mathrm{o}}] := \{a \in \{a\} : \mathsf{r}^{\mathrm{o}}(a) > 0\}$ of the opted r^{o}-factor is to be included in the action set $\{a\}$ allowed by the agent. The formula (5) implies $\mathrm{supp}[\mathsf{r}^{\mathrm{o}}] \subseteqq \mathrm{supp}[\mathsf{r}^{\mathrm{i}}]$. Thus, only the ideal r^{i}-factors

$$\mathsf{r}^{\mathrm{i}} \in \{\mathsf{r}^{\mathrm{i}}\} := \{\mathsf{r}^{\mathrm{i}} : \mathrm{supp}[\mathsf{r}^{\mathrm{i}}] = \{a\}\}, \tag{11}$$

keep actions in $\{a\}$ and exclude none. Thus, (11) is the generic constraint on r^{i}.

The next proposition, proved in [29], construct r^{io} in a subset of (11) that can approximate the optimum on (11) arbitrarily well.

Proposition 3 (Optimal r^{io}-Factor Meeting (11)): Let $\{\mathsf{r}^{\mathrm{i}}\}$ be given by $p > 1$

$$\{\mathsf{r}^{\mathrm{i}}\} := \{\mathsf{r}^{\mathrm{i}} : \mathrm{supp}[\mathsf{r}^{\mathrm{i}}] = \{a\} \text{ and } ||\mathsf{r}^{\mathrm{i}}||_p < \infty\}, \text{ while } |a| < \infty, \tag{12}$$

and let the assumptions of Prop. 2. hold. Then, the optimal ideal r^{io}-factor reads

$$\mathsf{r}^{\mathrm{io}}(a) \propto \chi_{\{a\}}(a) \exp[-\nu \mathsf{d}^{\mathrm{o}}(a)], \quad \nu := \frac{1}{p-1}, \tag{13}$$

$$\mathsf{d}^{\mathrm{o}}(a) := \int_{\{s\}} \mathsf{m}(s|a) \ln\left(\frac{\mathsf{m}(s|a)}{\mathsf{h}(s)\mathsf{m}^{\mathrm{io}}(s|a)}\right) \mathrm{d}s \overbrace{\leq}^{(10)} \mathsf{d}(a),$$

where χ denotes the set-indicator function.

The r^{io}-factor (13) is in (12) and thus it meets (11). For $p \to 1^+ \Leftrightarrow \nu \to \infty$, the set (12) fills arbitrarily tightly the set (11). Thus, the ideal rule (13) can be arbitrarily close to the optimum on (11).

The optimal ideal r^{io}-factor is uniquely given by m^{io} (symbolically, $\mathsf{r}^{\mathrm{io}} = \mathsf{r}^{\mathrm{io}}(\mathsf{m}^{\mathrm{io}})$) and by the opted $\nu > 1$, see (13). This allows us to meet the specific agent's wishes by opting $\mathsf{m}^{\mathrm{io}} \in \{\mathsf{m}^{\mathrm{i}}\}$ restricted by them. The following agent's generic wish

$$\boxed{\text{Reach ideal sets } \emptyset \neq \{s^{\mathrm{i}}\} \subseteq \{s\}, \emptyset \neq \{a^{\mathrm{i}}\} \subseteq \{a\}\,!} \tag{14}$$

is supported. Its adopted quantification guarantees that $\{\mathsf{c}^{\mathrm{i}}\text{-factors}\} \neq \emptyset$: (14) is taken as the wish to assign high probabilities to the given sets of ideal states $\{s^{\mathrm{i}}\}$ and of ideal actions $\{a^{\mathrm{i}}\}$ (14). The probabilities arise by closing the loop of the given, un-mutable, state-transition model with the optimal ideal decision rule $\mathsf{r}^{\mathrm{io}} = \mathsf{r}^{\mathrm{io}}(\mathsf{m}^{\mathrm{io}})$. This determines the optimum (13)

$$\mathsf{r}^{\mathrm{io}}(\mathsf{m}^{\mathrm{io}}) \in \operatorname*{Arg\,max}_{\mathsf{m}^{\mathrm{i}} \in \{\mathsf{m}^{\mathrm{i}}\}} \left[\int_{\{a\}} \rho(a)\mathsf{r}^{\mathrm{i}}(a)\,\mathrm{d}a\right] \tag{15}$$

$$:= \operatorname*{Arg\,max}_{\mathsf{m}^{\mathrm{i}} \in \{\mathsf{m}^{\mathrm{i}}\}} \left[\int_{\{a\}} \left[\int_{\{s\}} \chi_{\{s^{\mathrm{i}}\}}(s)\mathsf{m}(s|a)\,\mathrm{d}s + w\chi_{\{a^{\mathrm{i}}\}}(a)\right]\mathsf{r}^{\mathrm{i}}(a)\,\mathrm{d}a\right].$$

The weight $w \in \{w \geq 0\}$ assigns the importance of acting in $\{a^{\mathrm{i}}\} \subset \{a\}$ relatively to reaching $\{s^{\mathrm{i}}\} \subset \{s\}$. The sets $\{s^{\mathrm{i}}\}$, $\{a^{\mathrm{i}}\}$ is to be "reachable" on $\{a\}$ so that

$$\rho(a) > 0 \text{ on } \{a\}. \tag{16}$$

A few propositions, proved in [29], lead to a generic solution of (15) (and to the special "uniform" case, which is unpresented, but covered by Alg. 1). The solution uses

$$\bar{a} \in \operatorname*{Arg\,max}_{a \in \{a\}, \text{see (15)}} [\rho(a)] \tag{17}$$

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

Proposition 4 (m^{io} Meeting (15) for Generic $\mathsf{m}(s|a)$): Let $\mathsf{m}(s|a)$, $a \in \{a\}$, be a non-uniform pd on $\{s\}$ and conditions of Prop. 3 hold. Then, the m^{io}-factor meeting (15) reads

$$\mathsf{m}^{\mathrm{io}}(s|a) = \frac{\mathsf{m}(s|a)\exp[-\mathsf{e}(a)\mathsf{m}(s|a)]}{\int_{\{s\}} \mathsf{m}(s|a)\exp[-\mathsf{e}(a)\mathsf{m}(s|a)]\,\mathrm{d}s} \tag{18}$$

$$\text{well-defined for } |s| < \infty. \tag{19}$$

The real valued $\mathsf{e}(a)$ in (18) is the existing solution of the equation $\mathsf{L}(\mathsf{e}(a)) = \mathsf{R}(a)$, $a \in \{a\}$. The left- and right-hand sides of this equation are, see (17),

$$\mathsf{L}(\mathsf{e}(a)) := \mathsf{e}(a)\Lambda(a) + \ln\left[\int_{\{s\}} \mathsf{m}(s|a)\exp[-\mathsf{e}(a)\mathsf{m}(s|a)]\,\mathrm{d}s\right],$$

$$\Lambda(a) := \int_{\{s\}} \mathsf{m}^2(s|a)\,\mathrm{d}s > 0 \tag{20}$$

$$\mathsf{R}(a) := \int_{\{s\}} \mathsf{m}(s|a)\ln(\mathsf{h}(s))\,\mathrm{d}s + \mathsf{d}^{\mathrm{o}}(\bar{a}) + \ln\left[\frac{\rho(\bar{a})}{\rho(a)}\right] \geq 0.$$

C. Algorithm for Finite $|b|$

Alg. 1 summarises the theoretical results for closed-loops with a finite amount of behaviours. It makes the past state $\tilde{s} = s_{t-1}$ explicit.

The algorithm is well-applicable when using Bayesian estimation of unknown but time-invariant values of transition probabilities $\theta := (\theta_{s|a,\tilde{s}})_{s,\tilde{s} \in \{s\}, a \in \{a\}}$. The parametric model $\mathsf{m}(s_t|a_t, s_{t-1}, \theta) := \theta_{s_t|a_t,s_{t-1}}$ belongs to exponential family [2] and makes Dirichlet's prior pd self-reproducing. Its degrees of freedom counting the observed transitions $s_{t-1} = \tilde{s} \in \{s\}$, $a_t = a \in \{a\}$ to $s_t = s \in \{s\}$ form the sufficient statistic [25]. The randomised FPD actions allow to use the certainty-equivalent strategy that replaces unknown θ by its current point estimate. With a forgetting [31], the agent becomes adaptive. As usual, the certainty-equivalent strategy is implemented in the moving-horizon set-up: the strategy is re-designed whenever the parameter estimate is updated. The design horizon is to cover environment dynamics (length of its transients). Extensive references in [33] are the good starter for an updated insight into the used approximate strategy.

Algorithm 1 FPD with PE for Behaviours with a Finite Amount of Realisations

---

Inputs

✓ Sets of states $\{s\}$, actions $\{a\}$, ideal states $\{s^{\mathsf{i}}\} \subset \{s\}$ and ideal actions $\{a^{\mathsf{i}}\} \subset \{a\}$

✓ Relative weight $w \geq 0$ of $\{s^{\mathsf{i}}\}$, $\{a^{\mathsf{i}}\}$ *(15)*, Model $\mathsf{m}(s|a,\tilde{s})$, $s, \tilde{s} \in \{s\}, a \in \{a\}$

✓ Design horizon $|t|$, exploration controlling $\nu > 0$ & the function $\mathsf{h}(s) = 1$, $\forall s \in \{s\}$ *(4)*

Evaluation of h-independent variables

For $\tilde{s} \in \{s\}$ **do**

For $a \in \{a\}$ **do**

$\rho(a|\tilde{s}) = \sum_{s \in \{s^{\mathsf{i}}\}} \mathsf{m}(s|a,\tilde{s}) + \chi_{\{a^{\mathsf{i}}\}}(a)w$ *(15)*,

$\Lambda(a|\tilde{s}) = \sum_{s \in \{s\}} \mathsf{m}^2(s|a,\tilde{s})$ *(20)*

end $a \in \{a\}$

$\bar{a}(\tilde{s}) \in \mathrm{Arg}\max_{a \in \{a\}} \rho(a|\tilde{s})$, $\quad \bar{\rho}(\tilde{s}) = \rho(\bar{a}(\tilde{s})|\tilde{s})$

end $\tilde{s} \in \{s\}$

Design cycle for $t = |t|, |t|-1, \ldots, 1$

For $\tilde{s} \in \{s\}$ **do**

$\mathsf{d}^{\mathsf{o}}(\bar{a}(\tilde{s})) = \max\Big\{0,$
$\max_{a \in \{a\}} \Big[\sum_{s \in \{s\}} \mathsf{m}(s|a,\tilde{s}) \ln\Big[\frac{\rho(a|\tilde{s})}{\bar{\rho}(\tilde{s})\mathsf{h}(s)}\Big]\Big]\Big\}$

For $a \in \{a\}$ **do**

$\mathsf{d}^{\mathsf{o}}(a|\tilde{s}) = \mathsf{d}^{\mathsf{o}}(\bar{a}(\tilde{s})) + \ln\Big(\frac{\bar{\rho}(\tilde{s})}{\bar{\rho}(\tilde{s})}\Big)$

If $\mathsf{m}(s|a,\tilde{s})$ is not uniform

$\mathsf{R}(a|\tilde{s}) = \mathsf{d}^{\mathsf{o}}(a|\tilde{s}) + \sum_{s \in \{s\}} \mathsf{m}(s|a,\tilde{s}) \ln(\mathsf{h}(s))$ *(20)*

Find $\mathsf{e}(a|\tilde{s})$ in $\mathsf{R}(a|\tilde{s}) = \mathsf{e}(a|\tilde{s})\Lambda(a|\tilde{s})$
$+ \ln\big(\sum_{\{s\}} \mathsf{m}(s|a,\tilde{s}) \exp[-\mathsf{e}(a|\tilde{s})\mathsf{m}(s|a,\tilde{s})]\big)$

Set $\mathsf{m}^{\mathsf{io}}(s|a,\tilde{s}) \propto \mathsf{m}(s|a,\tilde{s}) \exp[-\mathsf{e}(a|\tilde{s})\mathsf{m}(s|a,\tilde{s})]$ *(18)*

else

Choose $\mathsf{o}(s)$ such that $\sum_{s \in \{s\}} \mathsf{o}(s) = 0$ [29]

Find $\mathsf{e}(a|\tilde{s})$ in $\ln\Big[\sum_{s \in \{s\}} \frac{\exp[-\mathsf{e}(a|\tilde{s})\mathsf{o}(s)]}{|s|}\Big] =$
$\mathsf{d}^{\mathsf{o}}(\bar{a}(\tilde{s}) + \frac{1}{|s|}\sum_{s \in \{s\}} \ln\Big[\frac{\mathsf{h}(s)\bar{\rho}(\tilde{s})}{\rho(a|\tilde{s})}\Big]$

Set $\mathsf{m}^{\mathsf{io}}(s|a) \propto \exp[-\mathsf{e}(a|\tilde{s})\mathsf{o}(s)]$.

end if on uniform m

$\mathsf{r}^{\mathsf{io}}(a|\tilde{s}) = \exp\big[-\nu \mathsf{d}^{\mathsf{o}}(a|\tilde{s})\big]$ *(13)*

end $a \in \{a\}$

$\mathsf{r}^{\mathsf{io}}(a|\tilde{s}) = \frac{\mathsf{r}^{\mathsf{io}}(a|\tilde{s})}{\sum_{\{a\}} \mathsf{r}^{\mathsf{io}}(a|\tilde{s})}$, $a \in \{a\}$ *(13)*

$\mathsf{n}(\tilde{s}) = \sum_{a \in \{a\}} \mathsf{r}^{\mathsf{io}}(a|\tilde{s}) \exp[-\mathsf{d}^{\mathsf{o}}(a|\tilde{s})]$,

$\mathsf{r}^{\mathsf{o}}(a|\tilde{s}) = \frac{\exp[-(\nu+1)\mathsf{d}^{\mathsf{o}}(a|\tilde{s})]}{\mathsf{n}(\tilde{s})}$, $a \in \{a\}$ *(4)*

end $\tilde{s} \in \{s\}$

$\mathsf{h}(s) = \mathsf{n}(s)$, $\forall s \in \{s\}$ *(4)*

end of the design cycle

Outputs All optimal ideal m^{io}, r^{io} and r^{o}-factors

III. Feedback via Meta-FPD with PE

The recalled DM with PE, referred as the basic DM, deals with two types of inputs:

✓ those directly describing the basic DM, which include:
- ► state $\{s\}$ and action $\{a\}$ sets;
- ► wishes-expressing ideal sets $\{s^{\mathsf{i}}\} \subset \{s\}$, $\{a^{\mathsf{i}}\} \subset \{a\}$;

✓ more technical, strategy-influencing, inputs that include:
- ► the weight $w \geq 0$ balancing the relative importance of ideal sets, see (15);
- ► the scalar $\nu > 1$ balancing exploitation with exploitation (duality, [17], [33]).

Fine modifications of ideal sets $\{s^{\mathsf{i}}\}$, $\{a^{\mathsf{i}}\}$ or the design horizon $|t|$ are other potential inputs of Alg. 1. For simplicity, the presentation focuses just on the pair w, ν. Its optimal choice depends on: ► the subjective agent's preferences; ► the agent's attitude to the basic DM; ► emotions, etc., all together on the agent's mental state. The dependence is complex and the mental state can hardly be directly measured and quantified. Thus, it is necessary to relate the optional inputs to the explicitly expressed user's satisfaction. The agent, referred to as the user in this case, is asked to judge the DM quality reached for various choices of inputs. This is the domain of classical PE [12] that often elicits preferences about a static DM and interactively queries the agent. Even advanced versions, represented by [7], become cumbersome in the targeted basic *dynamic* DM. This makes us adopt the next user-driven way that consists of formulating and solving an appropriate FPD meta-task.

The user assigns (satisfaction) marks, serving as the (meta) state $S_T \in \{S\}$, to the behaviour caused by the strategy, designed via Alg. 1 for trial values of the optional inputs (here, (w, ν)). Their changes A_T are as the (meta-)action. The actions are generated by (meta-)strategy gained by Alg. 1. It runs more slowly than the basic DM, $T \in \{T\} := \{\bar{T}, 2\bar{T}, \ldots, \} \subset \{t\}$ given by a step $\bar{T} > 1$. The applied zero-order holder keeps the latest agent's marking as the current state. This makes the agent quite free and allows the agent to stop the interactions according to its will.

This simple idea has to cope with the possible infinite regress, i.e. Alg. 1 at meta-level needs meta-inputs opted via a meta-PE, etc. Also, the curse of dimensionality [3] endangers applicability as the opted inputs are multiple and continuous-valued. The following way counteracts both obstacles.

The design horizon of the implemented certainty-equivalent strategy is to cover dominating dynamics of the closed-loop. This makes this horizon the natural smallest value of $\bar{T}$. Its multiples can be used if this rate is too high for the agent's marking. The use of a zero-order holder copes with the expected irregularity of the agent'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. II-C.

The choice of the ordinal scale of marks $\{S\} := \{1, \ldots, |S| := 5\}$ suffices for expressing "satisfaction degree". A rich, cross-domain, experience, e.g. in marketing [8] or in

European Credit and Accumulation System, confirms this. The mark $S = 1$ is taken as the best one, which unambiguously defines the ideal set $\{S^{\mathrm{i}}\} := \{1\}$.

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

$$\Delta w \in \{-\bar{w}, 0, \bar{w}\},\ \Delta\nu \in \{-\bar{\nu}, 0, \bar{\nu}\}, \quad \bar{w}, \bar{\nu} > 0. \quad (21)$$

Alg. 1 is to guarantee that opted inputs stay within their admissible ranges ($w \geq 0$, $\nu > 0$). The used simple clipping at boundaries of (21) seems to suffice. No other demands exist with respect to action. Thus, $\{A\} = \{A^{\mathrm{i}}\}$ and $W = 0$ (meta-twin to w in (15)). The last input to the meta-use of Alg. 1 is the counterpart of ν. This input cares about exploration that has to be stimulated at both levels. It makes no sense to choose a different value at the meta-level. Thus, ν is common at both levels: a slightly delayed value ν_{T-1} is at disposal when designing the new one.

The appearance of $\bar{T}$, $\bar{w}$, $\bar{\nu}$ demonstrates the danger of infinite regress. At present, it is cut by force and they are chosen heuristically. They, however, cover, the first step in a conceptual solution that: ▶ lets appear only meta-inputs that have a weak influence on results; ▶ tunes them via a universal adaptive minimisation of the mismodelling error [24].

IV. Experiments

Experiments primarily illustrate the presented theory. An extensive Monte Carlo study is under preparation and will be published elsewhere.

A. Common Simulation and Evaluation Options

a) Simulated environment: was chosen to be $15 \times 7 \times 15$ given by $|s| = 15$ and $|a| = 7$. It was created by learning the transition pd $\mathsf{p}(s_t|a_t, s_{t-1})$. 10^5 real values y_t stimulated by independently generated discrete actions in $\{a\} := \{1, \ldots, 7\}$ were used. The states $s_t \in \{s\} := \{1, \ldots, 15\}$ were gained via an affine mapping of discretised values of the real-valued y_t generated by ($y_0 = 0$)

$$y_t = 0.99 y_{t-1} + 0.05 a_t - 0.125 + 0.05\varepsilon_t.$$

There, ε_t is the white, zero-mean, normal noise. It has unit variance. The stationary expected level $s \approx 8$ for action $a \approx 4$ is interpreted as the zero "spent energy".

b) Experiments:: DM results without and with the user's control were compared. DM without the user control was the basic DM with no meta-level and wishes expressed by the ideal sets $\{s^{\mathrm{i}}\}$, $\{a^{\mathrm{i}}\}$ and by fixed options w, ν. DM with the user's control solved the basic DM supported by the second-layer implementing the solution of the meta-DM task as described in Sec. III. The DM with user's control gave the user the chance to express its satisfaction every ten steps, $\bar{T} = 10$. The satisfaction is quite subjective as it is demonstrated by presenting selected results for two different users, referred, 1st and 2nd user, respectively. Experimental conditions (see below) were set to make the results comparable. The users were informed about the key common conditions, i.e. the price paid for the respective action values, see Table I.

TABLE I
PRICE PAID FOR INDIVIDUAL ACTION VALUES

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

c) Experimental conditions:: Alg. 1 is used in the loop closed with the above environment.

Fixed options in all experiments were:

- the initial state $s_0 = 1$ and the seed of pseudo-random generator were reset to a common value in each experiments;
- the simulation length was 500 steps;
- sets of the ideal (desired) states $\{s^{\mathrm{i}}\}$ and actions $\{a^{\mathrm{i}}\}$ were fixed;
- the models (at both levels) were recursively estimated and the certainty-equivalent strategies with the receding horizon 100 were used;
- the prior statistics used in estimation determined uniform pds;
- $e = 1.2{*}\mathsf{ones}(|a|, |s|)$ initiated the search for $\mathsf{e}(.)$, Prop. 4;
- $p = 2 \Leftrightarrow \nu = \frac{1}{p-1} = 1$ was used in the cases without the user's control;
- the allowed changes of (w, ν) (21) were fixed to $\bar{w} = \bar{\nu} = 0.1$ in the cases with user's control.

The options distinguishing experiments were:

- user's control applied or not;
- the fixed values of w (15) in the cases without the user's control;
- the 1st or 2nd user expressed its satisfaction in the cases with the user's control.

B. Decision making without the user's control

1) Experiment 1.:

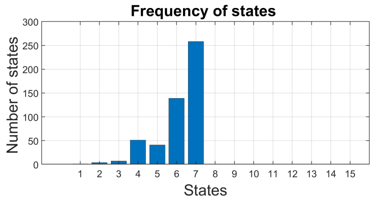
a) Experimental conditions: The user's wish is $\{s^{\mathrm{i}}\} = \{7\}$ an no extra wish is expressed on actions, $\{a^{\mathrm{i}}\} = \{a\}$.

b) Discussed results: The results are in Fig. 1. The desired state occurred the most often as we wanted and expected. All action values were realised with no extreme dominance of one value.

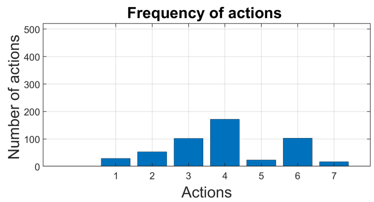
2) Experiment 2.:

a) Experimental conditions: The user's wish is $\{s^{\mathrm{i}}\} = \{7\}$ while requiring the actions to be in "zero energy" set $\{a^{\mathrm{i}}\} = \{4\}$. The weight value $w = 0.3$ (15) was fixed to express the latter wish.

b) Discussed results: The results are in Fig. 2. As it can be seen, the desired state has not occurred as often as in Exp. 1 due to the additional wish on actions. For $w = 0.3$, the desired action occurred the most often and the number of the desired action is much higher than in Exp. 1. This shows exactly what we wanted and expected.

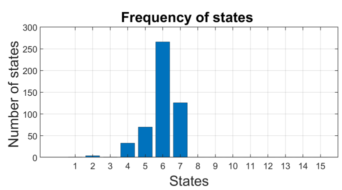


(a) States for $\{s^{\mathrm{i}}\} = \{7\}$, $\{a^{\mathrm{i}}\} = \{a\}$ $w = 0$

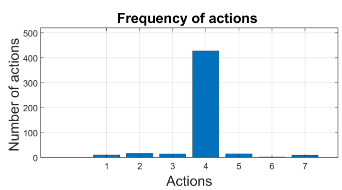


(b) Actions for $\{s^{\mathrm{i}}\} = \{7\}, \{a^{\mathrm{i}}\} = \{a\}$ $w = 0$

Fig. 1. Exp. 1: states and actions in DM without user's control and no wish on actions.



(a) States for $\{s^{\mathrm{i}}\} = \{7\}, \{a^{\mathrm{i}}\} = \{a\}$ $w = 0.3$



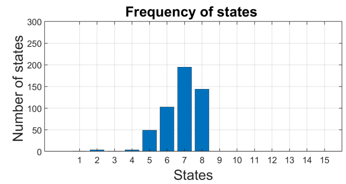
(b) Actions for $\{s^{\mathrm{i}}\} = \{7\}, \{a^{\mathrm{i}}\} = \{a\}$ $w = 0.3$

Fig. 2. Exp. 2: states and actions in DM without user's control and with a wish on actions.

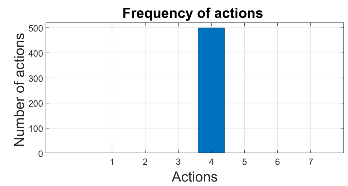
3) Experiment 3.:

a) Experimental conditions: The user's wish is $\{s^{\mathrm{i}}\} = \{7\}$ while requiring the actions to be in "zero energy" set $\{a^{\mathrm{i}}\} = \{4\}$ as in Exp. 2. The extreme weight $w = 10$ was tried.

b) Discussed results: The results are in Fig. 3. As expected the target state $\{s^{\mathrm{i}}\} = \{7\}$ is reached less often than in the previous case. The "harmonised" state $\{8\}$ is visited more often than before. The stress on the desired actions is surely too high. It is generally dangerous as the found strategy lacks the explorative capability. The same dangerous behaviour was observed for all $w \geq 1$.



(a) States for $\{s^{\mathrm{i}}\} = \{8\}, \{a^{\mathrm{i}}\} = \{4\}$ $w = 10$



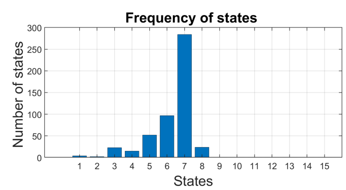
(b) Actions for $\{s^{\mathrm{i}}\} = \{8\}, \{a^{\mathrm{i}}\} = \{4\}$ $w = 10$

Fig. 3. Exp. 3: states and actions in DM without user's control and with a hard wish on actions

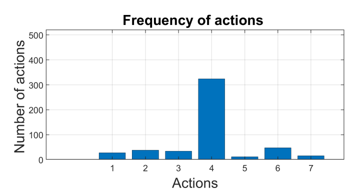
C. Decision making with the different user's control

1) Experiment 4.:

a) Experimental conditions: The user's wish was $\{s^{\mathrm{i}}\} = \{7\}$, $\{a^{\mathrm{i}}\} = \{4\}$. Neither the weight w nor ν were fixed and the 1^{st} user marked the seen closed-loop behaviour.



(a) States for $\{s^{\mathrm{i}}\} = \{7\}, \{a^{\mathrm{i}}\} = \{4\}$



(b) Actions for $\{s^{\mathrm{i}}\} = \{7\}, \{a^{\mathrm{i}}\} = \{4\}$

Fig. 4. Exp. 4: states and actions in DM with the 1^{st} user control

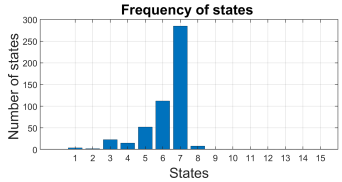
b) Discussed results: The results are in Fig. 4. As it can be seen the preferred state occurs most often. Compared to Exp. 1. without user's control, Experiment 4. gives better results.

Time courses of states, actions, weights w, exploration parameter ν and user's marks are in Figs. 6, 7. The corresponding discussion is there.

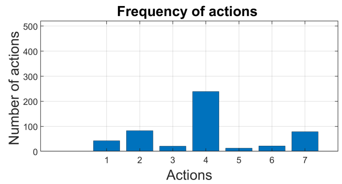
2) Experiment 5.:

a) Experimental conditions: The user's wish was $\{s^{\mathrm{i}}\} = \{7\}$, $\{a^{\mathrm{i}}\} = \{4\}$. Neither the weight w nor ν were fixed and the 2^{nd} user marked the seen closed-loop behaviour.

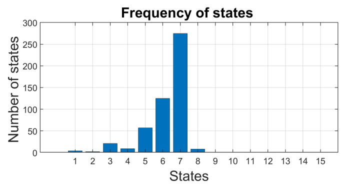
b) Discussed results: The results are in Fig. 5. They show how subjective individual preferences influence them.



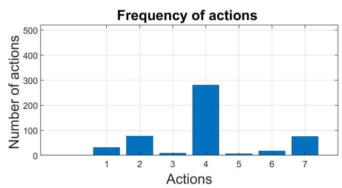
(a) States for $\{s^{\mathrm{i}}\} = \{7\}, \{a^{\mathrm{i}}\} = \{4\}$



(b) Actions for $\{s^{\mathrm{i}}\} = \{7\}, \{a^{\mathrm{i}}\} = \{4\}$



(c) States for $\{s^{\mathrm{i}}\} = \{7\}, \{a^{\mathrm{i}}\} = \{4\}$



(d) Actions for $\{s^{\mathrm{i}}\} = \{7\}, \{a^{\mathrm{i}}\} = \{4\}$

Fig. 5. Exp. 5: states and actions in DM with the 2nd user control

Objectively, this user paid a higher price, see Table II, but it did not get the desired state $\{s^{\mathrm{i}}\} = \{7\}$ as often as the 1st one. Time courses of states, actions, weights w, exploration parameter ν and user's marks are in Figs. 6, 7. The corresponding discussion is there.

D. Comparison of costs and responses in different experiments

Table II shows the price paid for actions in all experiments. It confirms expectations, including the desirable influence of users. The 1st user paid less than the 2nd one and less when no wish on actions is expressed. A (costly) expert's effort leading to a reasonable static compromise with $w = 0.3$ is possible.

a) Discussed results: Fig. 6 shows the evolution of the parameters and marks for both users. The 1st user was more consistent with his marking strategy. The marking by the 2nd user was more volatile: it gave almost every time a different mark. Fig. 6. Fig. 7 complements these trajectories by the time evolution of the states and actions. It can be seen that the marking strategy is more consistent for the 1st user. On the other hand, the parameter w stabilized faster for the 2nd user, but its paid price was higher, see Table II.

TABLE II
THE PRICE PAID FOR ACTIONS IN ALL EXPERIMENTS

Exp. no	Opted Parameters	Price
1	$w = 0, \ \nu = 1$	576
2	$w = 0.3, \ \nu = 1$	134
3	$w = 10, \ \nu = 1$	0
4	1st user	350
5	2nd user	610

V. Concluding Remarks

The paper advances the completion and quantification of preferences within the fully probabilistic design of decision strategies. The paper adds feedback that optimises optional inputs within the optimal ideal closed-loop model c^{io}. It needs as inputs: ▶ the set of allowed actions; ▶ specification of the desired state and actions sets; ▶ the on-line satisfaction marking by the user that judges behaviour improvements caused by changes of exploration option ν and of the scalar weights w balancing importance the ideal states and actions; ▶ online learnt and adapting the state-transition model.

The solution approaches the dreamt learning of preference [36]. It is worth stressing that the quantified preferences are both ambitious and realistic. Globally, it contributes to universal [21] and human-centric artificial intelligence [10].

The presented research is an open-ended story, which surely requires to deal with:

- ✓ collecting experience with our solution, initially, via extensive Monte Carlo studies;
- ✓ dimensionality curse connected with other wishes, say, balancing importance of state entries as needed in multi-attribute DM [1];
- ✓ counteracting the danger of infinite regress via [24] and thus challenging the claim that the quest for an absolute optimality is unrealistic [42];
- ✓ connection of the treated preference elicitation with an inattention level [43];
- ✓ specific application cases like [34]; etc.

These are definitely hard tasks requiring substantial intellectual effort. You are invited to expend yours.

References

[1] Azevedo, C., Zuben, F.V.: Learning to anticipate flexible choices in multiple criteria decision-making under uncertainty. IEEE Tran. on Cybernetics **46**(3), 778–791 (2016)

[2] Barndorff-Nielsen, O.: Information and Exponential Families in Statistical Theory. Wiley, N.Y. (1978)

[3] Bellman, R.: Adaptive Control Processes. Princeton U. Press, NJ (1961)

[4] Berger, J.: Statistical Decision Theory and Bayesian Analysis. Springer (1985)

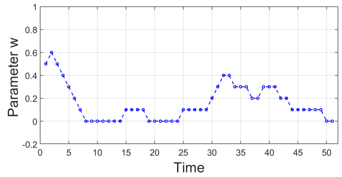
[5] Bertsekas, D.: Dynamic Programming and Optimal Control. Athena Sci. (2001)

[6] Bohlin, T.: Interactive System Identification: Prospects and Pitfalls. Springer (1991)

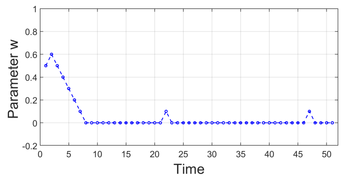
[7] Boutilier, C.: A POMDP formulation of preference elicitation problems. In: Proc. of the 18th National Conf. on AI, AAAI-2002. pp. 239–246. Edmonton, AB (2002)

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

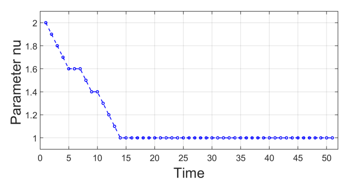
[9] Branke, J., et al: Efficient pairwise preference elicitation allowing for indifference. Computers & Oper. Res. **88**(Suppl. C), 175 – 186 (2017)



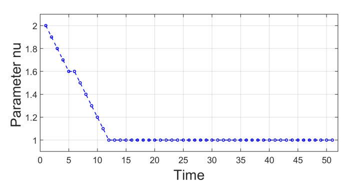
(a) Parameter w in time for Exp. 4.



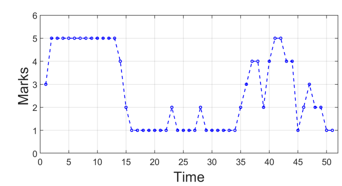
(b) Parameter w in time for Exp. 5.



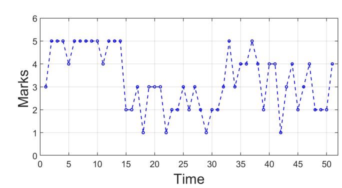
(c) Parameter ν in time for Exp. 4.



(d) Parameter ν in time for Exp. 5.



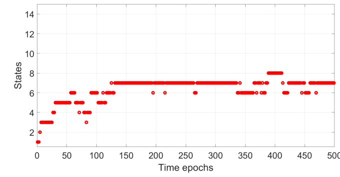
(e) Evolution of marks in time for Exp. 4.



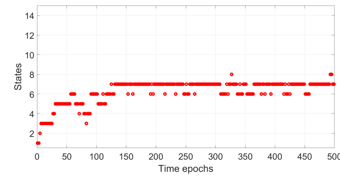
(f) Evolution of marks in time for Exp. 5.

Fig. 6. Evolution of the parameters w (15), ν (13) and user's marks with the 1^{st} and 2^{nd} user.

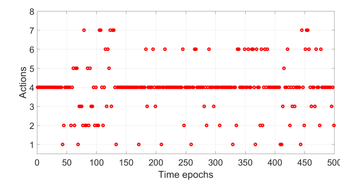
[10] Bryson, J., Theodorou, A.: How society can maintain human-centric AI. In: Toivonen, M., Saari, E. (eds.) Human-centered digitalization and services, pp. 305–324. Springer (2019)
[11] Bušič, A., Meyn, S.: Action-constrained Markov decision processes with Kullback-Leibler cost. In: Bubeck, S., Perchet, V., Rigollet, P. (eds.) Proc. of Machine Learning Research, vol. 75, pp. 1–14. MLR Press (2018)
[12] Chen, L., Pu, P.: Survey of preference elicitation methods. Tech. Rep.



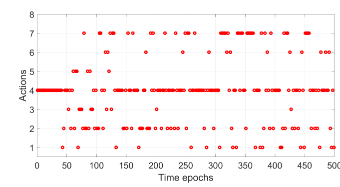
(a) States in Exp. 4.



(b) States in Exp. 5.



(c) Actions in Exp. 4.



(d) Actions in Exp. 5.

Fig. 7. Evolution of states and actions with 1^{st} and 2^{nd} user.

IC/2004/67, HCI Group Ecole Politechnique Federale de Lausanne, Switzerland (2004)
[13] Daee, P., Peltola, T., Soare, M., Kaski, S.: Knowledge elicitation via sequential probabilistic inference for high-dimensional prediction. Machine Learning **106**, 1599–1620 (2017)
[14] Drummond, J., Boutilier, C.: Preference elicitation and interview minimization in stable matchings. In: Proc. of 28th AAAI Conf. on AI. pp. 645 – 653 (2014)
[15] Dyer, J.S., Fishburn, P.C., Steuer, R.E., Wallenius, J., Zionts, S.: Multiple criteria decision making, multiattribute utility theory: The next ten years. Man. Sci. **38**(5), 645–654 (1992)
[16] Feinberg, E., Shwartz, A.: Handbook of Markov Decision Processes: Methods and Applications. Kluwer Academic Publishers (2002)
[17] Feldbaum, A.: Theory of dual control. Autom. Remote Control **22**, 3–19 (1961)
[18] Fishburn, P.: Nontransitive preferences in decision theory. Journal of Risk and Uncertainty **4**, 113–134 (1991)
[19] Guan, P., Raginsky, M., Willett, R.: Online Markov decision processes with Kullback Leibler control cost. IEEE Trans. on AC **59**(6), 1423–1438 (2014)
[20] Guy, T., Derakhshan, S.F., Štěch, J.: Lazy fully probabilistic design: Application potential. In: Belardinelli, F. (ed.) Multi-Agent Systems & Agreement Technologies (2018)
[21] Hutter, M.: Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability. Springer, Berlin, Heidelberg, N.Y. (2005)
[22] Kappen, H.: Linear theory for control of nonlinear stochastic systems. Physical review letters **95**(20), 200201 (2005)
[23] Kárný, M.: Axiomatisation of fully probabilistic design revisited. SCL, 104719 (2020). https://doi.org/10.1016/j.sysconle.2020.104719

[24] 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)

[25] Kárný, M., Böhm, J., Guy, T., Jirsa, L., Nagy, I., Nedoma, P., Tesař, L.: Optimized Bayesian Dynamic Advising: Theory and Algorithms. Springer, London, UK (2006)

[26] Kárný, M., Guy, T.: Fully probabilistic control design. SCL **55**, 259–265 (2006)

[27] 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)

[28] Kárný, M., Ruman, M.: Preference elicitation for Markov decision processes in fully probabilistic design set up. Applied Mathematics and COmputation (2021), submitted

[29] Kárný, M., Siváková, T.: Model-based preference quantification. IEEE Trans Cybernetics (2021), submitted

[30] Kracík, J., Kárný, M.: Merging of data knowledge in Bayesian estimation. In: Filipe, J., et al (eds.) Proc. of the 2nd Int. Conf. on Informatics in Control, Automation and Robotics. pp. 229–232. Barcelona (2005)

[31] Kulhavý, R., Zarrop, M.B.: On a general concept of forgetting. Int. J. of Control **58**(4), 905–924 (1993)

[32] Kullback, S., Leibler, R.: On information and sufficiency. Ann Math Stat **22**, 79–87 (1951)

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

[34] Perrault, A., Boutilier, C.: Experiential preference elicitation for autonomous heating and cooling systems. In: Proc. of the 18th Int. Conf. AAMAS '19. pp. 431–439 (2019)

[35] Peterka, V.: Bayesian system identification. In: Eykhoff, P. (ed.) Trends & Progress in System Identification, pp. 239–304. Perg. Press (1981)

[36] Pigozzi, G., , Tsoukiàs, A., , Viappiani, P.: Preferences in artificial intelligence. Annals of Mathematics and Artificial Intelligence **77**(3), 361–401 (2016)

[37] Quinn, A., Kárný, M., Guy, T.: Fully probabilistic design of hierarchical Bayesian models. Inf. Sci. **369**, 532–547 (2016)

[38] Quinn, A., Kárný, M., Guy, T.: Optimal design of priors constrained by external predictors. Int. J. Approximate Reasoning **84**, 150–158 (2017)

[39] Rao, M.: Measure Theory and Integration. J. Wiley (1987)

[40] Savage, L.: Foundations of Statistics. Wiley (1954)

[41] Shore, J., Johnson, R.: Axiomatic derivation of the principle of maximum entropy & the principle of minimum cross-entropy. IEEE Tran. on Inf. Th. **26**(1), 26–37 (1980)

[42] Simon, H.: Models of Bounded Rationality. MacMillan (1997)

[43] Sims, C.A.: Rational inattention: Beyond the linear-quadratic case. The Am. Econ. Rev. **96**(2), 158–163 (2006)

[44] Todorov, E.: Linearly-solvable Markov decision problems. In: Schölkopf, B., et al (eds.) Adv. in Neur. Inf. Proc., pp. 1369 – 1376. MIT Press (2006)