Modelling Human Optimal Seeking Behaviour During Evaluation of Process Models With Subjective Complexity

Ashish T. S. Ireddy[0000−0003−2964−4371], Leonid A. Beloglazovto, and Sergey V. Kovalchuk[0000−0001−8828−4615]

ITMO University, Saint Petersburg, Russia
{ireddy,kovalchuk}@itmo.ru

Abstract. Human–Artificial Intelligence (AI) collaboration in decision support scenarios is rapidly advancing. While rigorous effort is focused on improving AI models' solutions, deriving the experts' decision-making process remains a complex task. We present an experiment to model the optimal seeking behaviour of human experts as part of the human complexity state during information perceiving from process models of four real-world medical procedures. A targeted survey captures experts' perceived complexity evaluation using subjective metrics, understandability, correctness, and usability, in a free-to-explore grid world where process models of varying complexities are traversed and an optimal one is selected. We analyse expert trajectories to define three objective feature-based metrics: state discovery rate, state visitation entropy, and exploitation score, representing optimal seeking behaviour. Using Markov Decision Processes (MDP) and Maximum entropy Inverse Reinforcement Learning (MaxEnt IRL), we model exploration–exploitation during information perceiving and derive the underlying policy–reward function that reflects experts' subjective criteria for optimal model selection. Our results provide insights into human decision behaviour during optimal information seeking and assessment as an exploration-exploitation problem.

Keywords: Human optimal seeking · exploration-exploitation · process model evaluation · information perceiving · decision support systems

1 Introduction

The induction of AI agents and large language models (LLM) into professional sectors and daily life has allowed maximisation of efficiency and expanded outreach of knowledge. Information-intensive tasks can be significantly compressed while preserving reliability and coherence that would otherwise require an experienced domain specialist. However, AI agents still lack human-level awareness and thought process to consider real-world aspects such as rationality, situational relevance and contextual states [5]. While AI agent-based decision support systems (DSS) can assist and collaborate with human users in decision-making,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_26

the inclusion of excess or reduced information gives rise to vague or ambiguous solutions that may not align with the human users' expectations [2]. The exploration–exploitation construct provides a basis to balance information gathering and reward seeking. Yet, excessive exploration can yield stringent solutions while consuming enormous computational resources, whereas excessive exploitation can result in vague or biased solutions. [1] highlights the trade-off between performance and computational cost by considering the state of internal human planning before exploration as information seeking stage. [7] simulate a nuclear power plant accident to assess information seeking, integration, and diagnosis across individuals, tasks and trials. [8] and [6] observe that exploration can emerge from exploitation of objectives provided environments have recurring structures and agents can retain past experiences. The inclusion of human perceptional state and optimal solution-seeking behaviour into AI training can improve alignment with the user's expected solution pathways. We present an experiment to derive the optimal information-seeking behaviour of human experts in a decision-making scenario. Through a targeted survey, healthcare experts were invited to evaluate process models of four medical procedures ranging from minimum to maximum complexities. Experts freely explore a grid world environment and select a model representing optimal information via subjective metrics (understandability, correctness, usability) [5]. From expert trajectories, we model optimal seeking behaviour as an exploration–exploitation problem using three feature-based metrics: state discovery rate, state visitation entropy and exploration score; that characterize user exploration. Using MDPs, we simulate imitation learning and MaxEnt IRL to recover the underlying reward function and decision policy of human optimal selection. Our results provide a collective insight into deriving optimal seeking behaviour of experts in a decision making scenario. Further, the paper is structured as follows: Section 2 introduces the expert evaluation and methodology of modelling. Section 3 describes the interpretation of results from imitation learning and MaxEnt IRL, Section 4 investigates results after simulation. Section 5 is the conclusion.

2 Evaluation of process model complexity

In this section, we describe the experimental setup to collect human experts' evaluation of process model complexity via a targeted survey. We build on a prior experiment [3], where models describing four medical procedures were evaluated under limited dynamics. Here, we increase model complexity by varying *Activity rate* (AR) and *Path rate* (PR) (nodes and edges) spanning between (0, 0) to (100, 100), covering the entire state space.

Survey setup: We use a grid world environment for exploration of the state space by assigning process model control parameters (AR, PR) to (x, y) axes. We define a 11×11 grid (121 states) with steps at $[0, 10, \ldots, 100]$, selecting the corresponding process model at each (AR, PR). This resolution allows experts to reach maximal bounds of the state space without overly precise steps while preserving crucial complexity levels. The survey was distributed via personal

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_26

invitations to healthcare domain experts. On starting the survey, the experts were presented with an introduction to the evaluation task and anonymous demographic data collected. On starting the evaluation phase, (i) a 2D grid world, (ii) a control knob to select (AR, PR), (iii) a visualization of the corresponding process model, and (iv) background information on the medical process are shown. The control knob is initialized at state $(5, 5)$ to avoid bias or partial evaluation. Experts could freely explore the grid and select the process model deemed optimal. Changing the position of the control knob on the grid updates the visualized model. After optimal selection, the expert evaluated three subjective complexity metrics: **Understandability** (interpretability and legibility of the model), **Correctness** (accuracy of the presented information), and **Usability** (if information is sufficient to interpret the process). Figure 1 provides an overview of our experimental pipeline. At the time of this study, 16 responses were collected, with 10 fully usable. The survey continues to collect responses.

3 Modelling Human Optimal Seeking Behaviour

In this section, we introduce our approach to modelling human optimal seeking behaviour as part of human complexity states and notations used throughout the paper. A **grid world** environment G of dimensions $n \times n$ with grid **states** $s_i = (x_i, y_i) \in G_{n \times n}$. The **time spent** τ in each state is a state-space feature $\tau_i \in \mathbb{R}_{\geq 0}$, where each state is $\bar{s}_i = (x_i, y_i, \tau_i) \in G \ \{\mathbb{R}_{\geq 0}\}$. A set of **actions** in G is $a = \{\uparrow, \downarrow, \leftarrow, \rightarrow, \nwarrow, \nearrow, \searrow, \swarrow, stay\}$. **Displacement** is $a_i \equiv (\Delta x_i, \Delta y_i) \in \{-1, 0, 1\}^2$, where $\Delta x_i = x_i - x_{i-1}$ and $\Delta y_i = y_i - y_{i-1}$ for $i \geq 1$, with $a_0 = \{5, 5\}$ in all instances. Given a set of U **human expert responses**, the trajectory of state-transitions for an expert $u \in \{1, \ldots, U\}$ is $\mathcal{D}^{(u)} = \left\{ \left(s_i^{(u)}, a_i^{(u)}, \tau_i^{(u)} \right) \right\}_{i=0}^{J_u - 1}$, where J_u is the trajectory length for user u. We acquired 40 trajectories (10 experts across 4 datasets), each containing fields: *(x, y) trajectory*, timestamp, and *subjective evaluation*. We compute *movement metrics* describing expert behaviour characteristics such as displacement, final state, average time per state, etc.

Exploration - Exploitation: During model evaluation, the expected behaviour of an expert is traversal across min-max areas (information seeking) and comparison of complexity between (optimal seeking). We define *exploration* as the behaviour of new knowledge discovery by moving between states and *exploitation* as the behaviour of maximizing learned knowledge to select an optimal state with maximal reward. As each expert may have varying internal states (i.e. experience, expertise level, state of mind etc) we acknowledge the bias in users when reaching an optimal state i.e. efficiently (e.g. fewer steps, less discovery) or tediously (e.g. more interpretation). We term this as context [2], we will address this aspect in future works as it goes beyond the scope of this work. Hence, we introduce sliding windows ω that segment the expert's trajectory into sets of length $W \in \mathbb{N}$, with effective window length $L_i = \min(W, i + 1)$ and $\omega_i = \{i - L_t + 1, ..., i\}$, valid for $i \geq W - 1$. Given that a user has vis-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_26

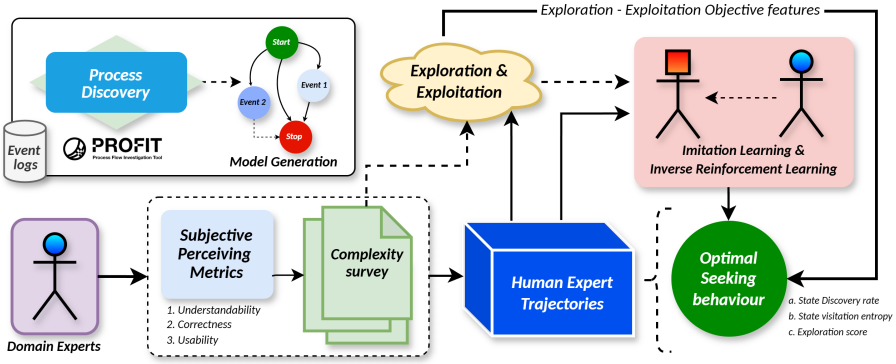

Fig. 1: Overview of our experimental setup to model human optimal seeking behaviour. Process models of varying complexities are evaluated by domain experts via a targeted survey. Exploration - exploitation behaviour during evaluation is formulated using objective feature metrics. Imitation learning and maximum entropy IRL recover underlying policy-reward function from expert trajectories.

ited states $V_{i-1} = \{s_0, s_1, ..., s_{i-1}\}$ before the i^{th} step, we introduce objective feature metrics that characterize the user's behaviour in the state space.

1. State discovery rate ($S^{discovery}$) measures states visited by the user for the first time (1 for new, 0 for revisits), where $S_W^{discovery}$ is the number of new states visited within W. The state revisit rate is defined as $S_W^{revisit}(i) = 1 - S_W^{discovery}(i)$. A High $S_W^{discovery}$ and $S_W^{revisit}$ score indicates high exploration and exploitation, respectively.

2. State visitation entropy: A Shannon entropy–based metrics that quantifies state visitation frequency and interpretation time along the user trajectory as *(A) Frequency based state visitation entropy* - derived from the empirical visitation distribution $p_{i,W}^{freq}(s) = \frac{C_{i,W}^{freq}(s)}{L_i}$ of the number of visits to grid states s as $C_{i,W}^{freq}(s) = |\{k \in W_i : s_k = s\}|$ in the sliding window ω_i with k states within, normalized over maximum entropy across the complete grid space as $(H_{i,W}^{freq})^{\text{norm}} = \frac{H_{i,W}^{freq}}{\log_2(|G|)}$. *(B) Time-based state visitation entropy* $\left(H_{t,W}^{\text{time}}\right)^{\text{norm}}$ follows the same principle as prior, but instead measures the time τ_k the user spends in each state s within window ω_i.

3. Exploration score: A heuristic measure of state-wise exploratory behaviour, defined as the sum of the state discovery rate and the normalized state visitation frequency over window W. This feature captures both the discovery of new states and the dispersion of exploration $E_{i,W} = S_W^{discovery}(i) + (H_{i,W}^{freq})^{norm}$.

4. Regime switch point: A metric to identify the change between exploration and exploitation regimes based on Cohen's d measure. Using expert trajectories in the grid world, we form a time series Q_i of exploration scores $E_{i,W}$ over T trajectories with i steps. We define $Q_i = E_{i,W}$ and split the series into two sections $n_1 = k + 1$ and $n_2 = T - (k + 1)$ where the assumed change of regime is at $k \in \{0, 1..., T - 2\}$ (i.e. we aim to compute the difference in exploration score between the two sections while varying the point of regime

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_26

inference between n_1 and n_2). For each k, we compute the mean exploration score $\mu(k) = \frac{1}{n}\sum_{i=0}^{k} x_i$, variance $\sigma^2(k)$, mean difference $\Delta(k) = \mu_1(k) - \mu_2(k)$ and the pooled standard deviation $s_{\text{pooled}}(k) = \sqrt{\frac{(n_1-1)\sigma_1^2(k)+(n_2-1)\sigma_2^2(k)}{n_1+n_2-2}}$ for each segment n_1 and n_2. We then compute Cohen's d as $d(k) = \frac{\Delta(k)}{s_{\text{pooled}}(k)}$ and maximize $d(k)$ for k steps $k^* = \arg\max_k d(k)$ reflecting the inferred point of change in exploration-exploitation regimes. Next, we extend this setup to select multiple points over sub-sections of the sequence, $\mathcal{K} = \{k_1, k_2, \ldots, k_m\}$ with $k_1 < k_2 < \cdots < k_m$, representing sub-interval switches over $I = [\ell, r]$ where $0 \le \ell < r \le T-1$ and $n_1 = k - \ell + 1,\ n_2 = r - k$. We compute the mean, variance, pooled standard deviation, mean difference, and Cohen's d over these sub-intervals as in $k^*(\ell, r) = \arg\max_{k\in\{\ell, r-1\}} d(\ell, r; k)$.

Imitation Learning and Inverse Reinforcement Learning Using experts' decision-making trajectories, and subjective evaluations we applied imitation learning (behavioural cloning) [9] to explore users' discovery paths when searching for optimal solutions, and maximum entropy inverse reinforcement learning (IRL) [10] to infer users' reward functions in the grid-world space. We define an MDP $\mathcal{M} = (S, A, T, \gamma)$ a 11×11 grid G with $|S| = 121$ states with starting state $s_0 = (5, 5)$. A set A containing 9 actions with state transition function $T(s, a) = \pi_s\big(s + \Delta(a)\big)$, mapping state–action pairs through policy π. Expert demonstrations trajectories as $E^i_{traj} = \{(s_0, a_0), (s_1, a_1), \ldots, (s_i, a_i)\}$ for user i, with δ_i the time the user spends in state i. We extend the exploration–exploitation metrics: state discovery rate $S^{discovery}$, state visitation entropies $H^{\text{time}}_{t,W}$ and $H^{\text{freq}}_{t,W}$, visit distribution $C^{freq}_{i,W}(s)$, and trajectory direction geometry θ, computed from the step vectors and the agent's movement orientation from Action space A. Behavioural Cloning attempts to learns a policy $\hat{\pi}(a \mid \phi(s_i))$ from state–action pairs (s_t, a_t) in a supervised manner, using feature vectors $\phi(s_i)$ $\phi(s_i) = [x_{(s_i)}, y_{(s_i)}, C^{freq}_{i,W}(s),\ \delta_{(s_i)}, \theta_{s_i}, H^{\text{norm}}_{t,W}, S^{discovery}_{s(i)}]$. Using a logistic regression function, we simulate agent traversal for 2000 iterations. At each iteration we infer $a_i = action(s_i, s_{i+1})$ where $s_{i+1} = T(s_i, a_i)$ and estimate parameters by maximizing entropy: $\phi^* = \arg\max_\Theta \sum_{(\phi_i, a_i)\in E^i_{traj}} \log \pi_\Theta(a_i \mid \phi_i)$; where E^i_{traj} is sampled from individual or group user trajectories. We use the grid world MDP to implement MaxEnt IRL to find an optimal policy π^* maximizing reward function R. We define a linear reward function $R_\omega(s) = \omega^\mathsf{T} f(s)$ based on exploration variables AR–PR (x, y) and descriptive QUAD metrics (replay fitness, precision, generalization, simplicity) acquired for each process model [4], where $\omega = [w_0, w_{\text{fit}}, w_{\text{prec}}, w_{\text{gen}}, w_{\text{simp}}, w_x, w_y]^\mathsf{T}$, $F \in \mathbb{R}^{121}$ is the feature matrix for transitions, and $f(s)$ is a row of F. Next, we fit the reward weights ω from $R_\omega(s)$ to obtain the assumed optimal policy π^*_ω depicting reward behaviour.

4 Case study: Investigating Information Seeking

In this section, we investigate the results acquired from expert trajectories during optimal seeking. We compute objective feature metrics for all expert trajectories across four datasets to assess exploration-exploitation behaviour, resulting

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_26

in a large set of observations. Each user had unique information-seeking and reward alignment approaches for optimal model selection (i.e. discovery of the grid space, complexity bounds, trajectory steps etc). As our work is focused on deriving optimal seeking behaviour, we aim to address expert personalization, dataset context, descriptive metrics and trajectory behaviour in future works. To ensure a consistent scope of results, we select one expert's evaluation of process models describing a hospital billing process and collectively analyse the behaviour across all other responses with a sliding window of $\omega = 25$. Figure 2 visualizes the trajectory followed by the expert during the evaluation phase. The colour density represents the interpretation time spent in each cell. The expert discovered $\sim 60\%$ of the process complexity space while spending significant interpretation time in spaces with higher (AR, PR), i.e. more complex models. Figure 2 plots the objective feature metrics over trajectory steps. At the start of the exploration regime, the $S^{discovery}$ score is high as the state space is unexplored. With visitation to new states, the score gradually decays, while $S^{revisit}$ proportionally increases. At certain steps, we observe that discovery scores remain relatively constant or marginally increased; we identify this phase as exploitation, where the user compares already explored models to find an optimal. This behaviour is also reflected in state visitation entropy, where the expert is already in an exploitation regime. The lines n_1 and n_2 in plots 2 reflect instances of the expert switching from exploration to exploitation, where explored states are revisited and compared for optimal information. Next, we simulated the results of increasing ω from 5 to 100 in steps of 10 and observed that once ω increases over 50% of the trajectory length, the occurrence of at least one switch point is low, whereas having an extremely small window size reflects marginally small metric scores to be flagged as a change in exploration-exploitation regimes. Hence, there is also a need to select an ideal ω such that it is constrained by the trajectory length and knowledge gained. We also acknowledge that in expert trajectories, the grid area with low complexity models was left unexplored, this can be attributed to the nature of the data presented or the use case of the model, i.e. having an abstract process model is not logical to describe the medical process. This can be considered an underlying latent state as part of human perception.

Further, we acquire the results of imitation learning and MaxEnt IRL over the grid world MDP shown in Figure 3A and 3B. We simulate the exploratory behaviour for a different expert when evaluating process models describing internal operations of a hospital. The expert here has visited both extremes of the grid space and discovered the minima and maxima models. We observe the learning agent's path to follow similar behaviour during exploitation as the expert did not visit the low complexity cells after first time exploration. The IRL reward function reflects the internal optimal reward function assumed by the user's exploration and exploitation. It reveals high concurrent rewards when exploiting cells with higher QUAD scores and greater process model complexity. It is worth noting that the reward at cell $(10, 10)$ is lower than the region of $(9, 9)$ to $(6, 6)$. This reflects both the user's behaviour of exploitation at the maximum complexity (high levels of information that may be deemed overcomplicated)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_26

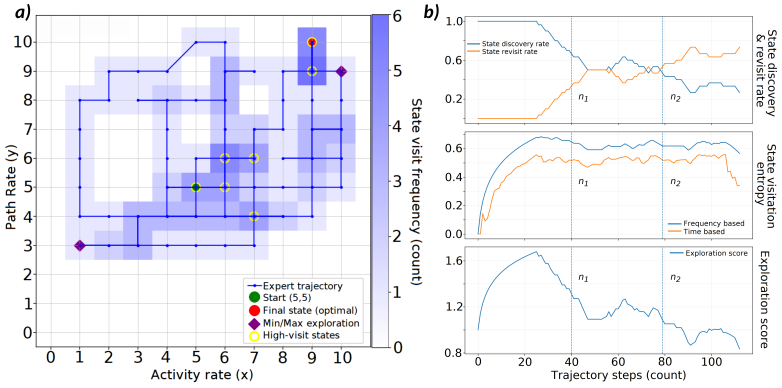

Fig. 2: Overview of exploration-exploitation behaviour of an expert evaluating process models describing a hospital billing process. (A) The expert trajectory in the grid world space where colour density represents the interpretation time spent in the cell; (B) The objective feature metrics of the expert plotted over trajectory steps at a sliding window size ($\omega = 25$), (n_1, n_2) represent the regime change from exploration to exploitation.

and the descriptive metrics that indicate the model is not usable. Further, we confirm this behaviour by comparing the time spent by the expert within these states, reflecting exploitation. This trend was also observed in other datasets and experts, where the agent differentiated between the frequency of visitations and the interpretation time per state to replicate the optimal seeking behaviour of the expert during exploitation

Fig. 3: (A): Expert trajectory (in blue) and imitation learning agent behaviour (in orange) when evaluating process models describing internal operations of a hospital (B): The recovered reward function from Maximum entropy IRL

5 Conclusion

Our work presents an approach to modelling human optimal seeking behaviour in a grid world scenario during the evaluation of process model complexity. Expert trajectories from a targeted survey were collected and exploration-exploitation

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_26

behaviour was derived via three objective feature metrics State discovery rate, state visitation entropy and exploration score. Using a sliding window, we characterise information-gathering and reward-seeking regimes to identify the change from exploration to exploitation in the human user's perceptional state. We implement imitation learning and maximum entropy inverse reinforcement learning to replicate the exploration-exploitation behaviour and to recover the underlying reward function of the human expert during evaluation. Our results provide a crucial step towards modelling the human solution-seeking behaviour in decision-making scenarios. In future works, we plan to extend the study by continuing analysis with the inclusion of latent complexity states, context, personalization, descriptive and subjective metrics, to model human behaviour with online training of AI agents with feedback during interaction. More details on the background, implementation and elaboration of achieved results is available at https://github.com/maddoxx02/ICCS-2026-Modelling-Optimal-Seeking.

Acknowledgements. The research was supported by The Russian Science Foundation, agreement №24-11-00272, https://rscf.ru/project/24-11-00272/.

References

1. Callaway, F., Van Opheusden, B., Gul, S., Das, P., Krueger, P., Lieder, F., Griffiths, T.: Human planning as optimal information seeking. Manuscript in prep. (2021)
2. Ireddy, A.T.S., Kovalchuk, S.V.: Analysis of internal and external context in clinical decision scenarios with expert feedback. In: International Conference on Intelligent Information Technologies for Industry. pp. 301–313. Springer (2025)
3. Ireddy, A.T., Ionov, M.V., Beloglazov, L.A., Zatsepina, E.A., Kovalchuk, S.V.: Evaluating perceived complexity of process models from a targeted survey of healthcare domain specialists. In: International Conference on Mathematical Modeling and Supercomputer Technologies. pp. 43–58. Springer (2024)
4. Ireddy, A.T., Kovalchuk, S.V.: An experimental outlook on quality metrics for process modelling: a systematic review & meta analysis. Algorithms **16**(6), 295 (2023)
5. Kovalchuk, S., Ireddy, A.T.S.: Prediction of users perceptional state for human-centric decision support systems in complex domains through implicit cognitive state modeling. In: Proceedings of the Annual Meeting of the Cognitive Science Society. vol. 46 (2024)
6. Kulich, M., Krajník, T., et al.: To explore or to exploit? learning humans' behaviour to maximize interactions with them. In: International Workshop on Modelling & Simulation for Autonomous Systems. pp. 48–63. Springer (2016)
7. Lyu, X., Li, Z.: Predictors for human performance in information seeking, information integration, and overall process in diagnostic tasks. International Journal of Human–Computer Interaction **35**(19), 1831–1841 (2019)
8. Rentschler, M., Roberts, J.: Exploitation is all you need... for exploration. arXiv preprint arXiv:2508.01287 (2025)
9. Torabi, F., Warnell, G., Stone, P.: Behavioral cloning from observation. arXiv preprint arXiv:1805.01954 (2018)
10. Ziebart, B.D., Maas, A.L., Bagnell, J.A., et al.: Maximum entropy inverse reinforcement learning. In: Aaai. vol. 8, pp. 1433–1438. Chicago, IL, USA (2008)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_26