PoETree: Interpretable Policy Learning with Adaptive Decision Trees

Anonymous authors
Paper under double-blind review

Abstract

Building models of human decision-making from observed behaviour is critical to better understand and support real-world policies such as clinical practice. This marks a paradigm shift in policy learning, from performance-focused goals to the challenge of faithfully representing and explaining the demonstrated decision-making process. Policy Extraction through decision Trees (PoETree) is a novel framework for interpretable policy learning, compatible with fully-offline and partially observable clinical decision environments. Our probabilistic, fully-differentiable tree architecture is grown incrementally during optimization to adapt its complexity to the modelling task, and integrates history representation learning by recurrence. The resulting decision tree policies, adapting over time with patient information, outperform state-of-the-art policy learning methods on real and synthetic medical datasets, both in terms of understanding, quantifying and evaluating observed behaviour as well as in accurately replicating it – with potential to improve future decision support systems.

1 Introduction

Different approaches to integrating, analysing and acting upon clinical information leads to wide, unwarranted variation in medical practice across regions and institutions (McKinlay et al., 2007; Westert et al., 2018). There is therefore a demonstrated potential for machine learning to support the clinical decision-making process, with algorithms successfully deployed for oncology prognosis (Beck et al., 2011; Esteva et al., 2017), retinopathy detection (Gulshan et al., 2016) or radiotherapy treatment design (Valdes et al., 2017).

Performance and patient safety have also been shown to improve when incorporating medical domain knowledge, benefiting from both computational power and scientific expertise. Domain knowledge was critical to the development of robotic-assisted surgery, guided by surgeons through imitation learning (Kassahun et al., 2015). Still, these support systems focus on effectively replacing physicians with autonomous agents, trained to optimise patient outcomes or diagnosis accuracy. With limited design concern for end-users, these algorithms fall short on interpretability, leading to mistrust from the medical community (Laï et al., 2020; The Royal Society, 2017).

Instead, our work aims to better describe and *understand the decision-making process* by observing physician behaviour, in a form of epistemology (Jarrett et al., 2021). Domain expertise from current practice can both be captured and explicitly incorporated into such policy models, thus enabling quantitative, objective comparisons of decision-making pathways. In turn, this could lead to novel guidelines and educational materials. Finally, as an end-goal, these transparent models of behaviour can form the building blocks of future decision support systems (Li et al., 2015), designed in partnership with clinicians for more reliable validation, better trust, and widespread acceptance.

The associated machine learning challenge, therefore, is to determine what action to take given previous patient observations, and thus describe an interpretable decision-making policy: *Given new patient observations and their previous medical history, what action do physicians take and why?* Modelling clinical diagnostic and treatment policies particularly challenging, with observational data involving missing or confounding factors and unknown evolution dynamics, which we wish to then translate into a transparent representation of decision-making. State-of-the-art solutions for uncovering demonstrated policies, such as inverse reinforcement learning (Ng & Russell, 2000) or imitation learning (Bain & Sammut, 1996; Piot et al., 2014; Ho & Ermon, 2016), still fall short on interpretability – mainly proposing black-box models as behaviour representations.

In contrast, *decision trees* models are highly familiar to the medical community, with many clinical guidelines having this form (Qaseem, 2012; Chou et al., 2007; McCreery & Truelove, 1991; Burd & Sonnenberg, 2002; Burch et al., 2012). Cognitive research suggests human decision-making does follow a similar hierarchical, branching process (Zylberberg et al., 2017), making decision trees straightforward to understand. As a result, we set out to represent clinical policies through decision trees, evolving over time as patient information is accumulated.

Contributions The main contributions of this work are as follows: (i) A novel framework for behavioural cloning on tabular data in the form of incrementally grown, probabilistic decision trees, which adapt their complexity to the task at hand. (ii) An interpretable model for representation learning over time-series with actions called recurrent decision trees. (iii) An evaluation of our algorithm on real and synthetic medical datasets, demonstrating effectiveness both in understanding and replicating observed behaviour. In particular, our work stands out for its expressive power in recovering ground-truth policies – disambiguating factors that are unidentifiable with related benchmarks; for its integration of domain knowledge through inductive bias; and for its formalisation and quantification of subjective behavioural analyses.

2 Problem Formalism

Our goal is to learn an *interpretable representation of a demonstrated decision-making process* to understand agent behaviour. The clinical context requires *offline learning* using observational data only, as experimenting with policies would be both unethical and impractical. Finally, as previous observations and actions affect treatment choice, we are concerned with a *partially-observable* Markov Decision Process (PO-MDP).

We assume access to a dataset of m demonstrated patient trajectories in discrete time $\mathcal{D} = \{(z_1^i, a_1^i, \ldots, z_{\tau_i}^i, a_{\tau_i}^i)\}_{i=1}^m$, where τ_i is the length of trajectory i. We drop index i unless explicitly required, to denote a generic draw from the population. At timestep t, the physician-agent observes patient features, denoted by random variable $Z_t \in \mathcal{Z}$, and chooses treatment or diagnostic $A_t \in \mathcal{A}$, where $\mathcal{A}$ is a finite set of actions. Let z_t and a_t denote realisations of these random variables. The environment's reward function R, state space $\mathcal{S}$, transition dynamics and observation function are unknown. As we are only concerned with recovering experts' thought-process, we can assume full access to their observations.

Let $h_t = (z_1, a_1, ...z_{t-1}, a_{t-1}) = (z_{1:t-1}, a_{1:t-1}) \in \mathcal{H}$ denote a realisation of random variable $H_t \in \mathcal{H}$, the patient history of observations and actions at timestep t, or information accessible prior to observing a new set of features z_t. Following the PO-MDP formalism, information in $\{h_t, z_t\}$ can be interpreted to form a belief over the true patient state s_t, inaccessible to the learning agent. In contrast to recent work on partially-observable environments (Sun et al., 2017; Choi & Kim, 2011; Makino & Takeuchi, 2012; Hüyük et al., 2021), our goal is to highlight how newly acquired z_t conditions action a_t, to ensure decisions are based on interpretable, meaningful variables. As a result, observed behaviour must be represented as a mapping *from observation to action space*.

A stationary stochastic *decision-making policy* is a mapping $\pi : \mathcal{Z} \times \mathcal{H} \times \mathcal{A} \to [0, 1]$, where $\pi(a_t|z_t, h_t)$ encodes the probability of choosing an action a_t given all available patient information, including patient history h_t and the latest patient observation z_t, such that $\sum_{a_t \in \mathcal{A}} \pi(a_t|z_t, h_t) = 1$ at every timestep t. We assume that trajectories $\mathcal{D}$ are generated by an agent following an expert policy π_E, such that $a_t \sim \pi_E(\cdot|z_t, h_t)$. The goal of our work is to find an *interpretable* representation of this adaptive policy π which matches and explains the behaviour of π_E.

3 Interpretable policy learning with decision trees

3.1 Soft decision trees

We propose a policy learning model in the form of a soft decision tree, a fixed tree structure with binary nodes and probabilistic decision boundaries (Irsoy et al., 2012). This supervised learning framework can be used for behavioural cloning (Bain & Sammut, 1996; Piot et al., 2014) by setting input variables $\mathbf{x} = z_t$ (for a fully-observable environment), or in the non-Markovian case: $\mathbf{x} = z_{1:t} = \{h_t, z_t\}$. Completing the policy mapping, targets are $\mathbf{y} = a_t$.

Inner nodes The path followed by input $\mathbf{x} \in \mathbb{R}^D$ is determined by gating functions $p_{gate}(\mathbf{x})$ at each non-leaf node, encoding the probability of taking the rightmost branch as in Figure 1a. The

path probability of each node l, denoted $P^l(\mathbf{x})$, is the product of gating probabilities leading to this node from the tree root. We choose a multidimensional sigmoid over normalised features as gating function (Frosst & Hinton, 2017):

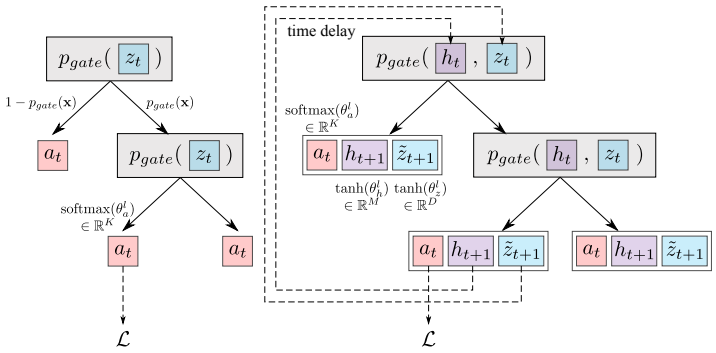
$$p_{gate}(\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + b) = \sigma\left(\sum_{i=1}^{D} w_i x_i + b\right) \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^D$ and $b \in \mathbb{R}$ are trainable parameters for each inner node and σ is the sigmoid function. The resulting challenge is how to make this splitting interpretable: Silva et al. (2020) suggest to select the largest component of $\mathbf{w}$ at each node to obtain an *axis-aligned* threshold over a single input dimension as detailed in Appendix C. Model simplification can also be made more robust by encouraging sparsity in $\mathbf{w}$ through L1 regularisation.

Leaf nodes Leaf nodes in our tree architecture define a fixed probability distribution over K categorical output classes (for a policy choosing between treatment or diagnostic actions). Leaf l encodes relative activation values $a_t^l = \text{softmax}(\theta_a^l)$, where $\theta_a^l \in \mathbb{R}^K$ are learned leaf parameters. Discrete or continuous action types could also be modelled by passing leaf parameters through a hyperbolic tangent function. For ease of computation and interpretation, the output of the maximum-probability leaf $a_t^{l_{max}}$, where $l_{max} = \arg\max_l P^l(\mathbf{x})$, is taken as the tree's overall classification decision, following Frosst & Hinton (2017).

3.2 Recurrent decision trees

Policy learning in POMDPs must account for previously acquired information in decision-making by extracting a compact representation of patient history h_t. We propose *recurrent decision trees* (RDTs) to learn similar mappings to recurrent neural networks (RNNs) with the interpretability and probabilistic framework afforded by soft decision trees. Representation and policy learning are jointly optimised such that, for a given history h_t and new observation z_t as tree inputs, decision paths lead to leaves that output *both* the predicted action a_t as well as the history embedding for the next timestep h_{t+1}. Shown in Figure 1b, the tree architecture is thus modified for leaves to output both history and actions by assigning an additional set of parameters $\theta_h^l \in \mathbb{R}^M$, where M is the history dimensionality, to each leaf l; history output values are computed as $h_{t+1}^l = \tanh(\theta_h^l)$ (with more complex recurrence models in Appendix B). A single structure is used at all timesteps: unfolded over time, our recurrent tree can be viewed as a sequence of cascaded trees (Ding et al., 2021) with shared parameters and structure. These differentiable structures can be optimised by backpropagation through time as for an RNN.



(a) DT policy. (b) RDT policy.

Figure 1: **Decision tree policy structure, with different history extraction models**. Loss $\mathcal{L}$ is given by $L + L_{split} + L_{\tilde{z}}$.

Expected patient outcomes Our multi-output model can also be extended to predict patient outcomes or observations at the next timestep $\tilde{z}_{t+1}$ by passing an additional set of leaf parameters θ_z^l through a hyperbolic tangent. This is valuable (i) for providing *decision-making explanations* in terms of treatment effects under action a_t (Yau et al., 2020); (ii) for highlighting *unexpected patient evolution* through differences in outcome $(\tilde{z}_t - z_t)$; (iii) finally, predicted features $\tilde{z}_t$ can be used as proxies for new observations z_t when approximating multidimensional trees as axis-aligned ones, thus *improving the quality of threshold values* as detailed in Appendix C.

3.3 Tree growth and optimisation

Optimisation objective Behavioural cloning exploits supervised learning loss functions to recover the mapping from observations to actions, as demonstrated by physician agents. For a tree topology

$\mathbb{T}$ with parameters Θ, our loss function combines the cross-entropy between one-hot encoded target $a_t \in \{0,1\}^K$ and each leaf's output, weighted by its respective path probability under input $\{h_t, z_t\}$:

$$L(\{h_t, z_t, a_t\}; \mathbb{T}, \Theta) = -\sum_{l \in LeafNodes} P^l(h_t, z_t) \sum_k a_{t,k} \log Q_k^l \tag{2}$$

where $P^l(h_t, z_t)$ is the path probability of leaf l, and Q_k^l is its output probability for class k. For discrete and continuous actions, the cross-entropy is replaced with the mean-squared error. To improve optimisation, Frosst & Hinton (2017) also encourage the model to split data evenly across tree inner nodes through an additional regularisation term L_{split}. Finally, regularisation term $L_{\tilde{z}}$ is designed to learn outcomes $\tilde{z}_t$ as expected by the acting physician. In addition to minimising prediction error on z_t, we ensure predicted observations lead to similar action choices as true ones, such that the policy is consistent between timesteps:

$$L_{\tilde{z}}(\{h_t, z_t, a_t\}; \mathbb{T}, \Theta) = \delta_1 \left\| z_t - \tilde{z}_t \right\|^2 + \delta_2 \, D_{\text{KL}} \left(\pi(\cdot|h_t, z_t) \| \pi(\cdot|h_t, \tilde{z}_t) \right) \tag{3}$$

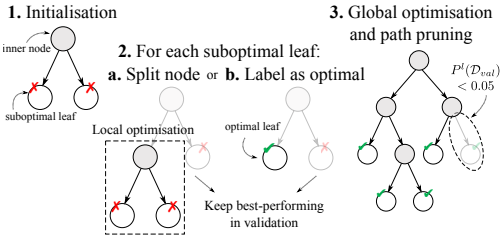
where $\{\delta_1, \delta_2\}$ are tunable weights, trading-off fidelity to true outcomes and fidelity to demonstrated behaviour. For a set of training trajectories $\mathcal{D}$, our behavioural cloning optimisation objective becomes:

$$\underset{\mathbb{T},\Theta}{\arg\min} \; \left[L + L_{split} + L_{\tilde{z}} \right] (\mathcal{D}) \tag{4}$$

The probabilistic nature of our model choice makes it fully-differentiable, allowing optimisation of parameters Θ through stochastic gradient descent and backpropagation through the structure for a fixed topology $\mathbb{T}$. Optimising with respect to $\mathbb{T}$, however, requires a more involved algorithm.

Algorithm 1: **Tree growth optimisation**.

1. Initialise $\mathbb{T}$ to inner node with two suboptimal leaves;
Optimise Θ via gradient descent of loss $\mathcal{L}$;
while $\exists$ *suboptimal leaves in* $\mathbb{T}$ **do**
  Split suboptimal leaf into inner node with two leaves;
  2a. Locally optimise Θ via gradient descent of $\mathcal{L}$;
  if *validation performance improves* **then** retain split;
  else 2b. Retain leaf as optimal;
3. Globally optimise Θ via gradient descent of $\mathcal{L}$;
Prune branches in $\mathbb{T}$ with low path probability $P^l(D_{val})$.



Tree growth Following the work of Irsoy et al. (2012) and Tanno et al. (2019), our decision tree architecture is incrementally grown and optimised during the training process, allowing the model to independently adapt its complexity to the task at hand. Algorithm 1 summarises the growth and optimisation procedure. Starting from an optimised shallow structure, tree growth is carried out by sequentially splitting each leaf node in breadth-first order, up to pre-determined depth, and locally optimising new parameters – the split is accepted if validation performance is improved. After a final global optimisation, branches with a path probability below a certain threshold on the validation set are pruned. Further details on the tree growth procedure are given in Appendix B.

Complexity Policy training can be computationally complex due to the multiple local tree optimisations involved and significant number of parameters in comparison to other interpretable models (Hüyük et al., 2021). Still, at test time, computing tree path probabilities for a given input is inexpensive, which is advantageous to allow fast decision support. A detailed comparison of runtimes involved with our method and related work is proposed in Appendix D.

4 Related Work

In this section, we compare related work on interpretable policy learning, and contrast them in terms of our key desiderata. In the following paragraphs we discuss the literature in depth, but key aspects and an overview are summarised in in Table 1.

Policy learning While sequential decision-making problems are traditionally addressed through reinforcement learning, this framework is inapplicable to our policy learning goal due to the unavailability of a reward signal R from the environment. Instead, we focus on the inverse task of replicating the demonstrated behaviour of an agent, known as *imitation learning* (IL), which assumes no access to a reward signal. Common approaches include behavioural cloning (BC) where this task is reduced to supervised learning, directly mapping states to actions (Bain & Sammut,

Table 1: **Related work.** Interpretability comparison to policy learning methods.

Related work	Offline Learning	Partial Observability	Transparency	Modelling Assumptions
BC-IL (Bain & Sammut, 1996)	✓	✗	✓	$\mathcal{S} = \mathcal{Z}$
PO-BC-IL (Sun et al., 2017)	✓	✓	✗	$\mathcal{A}, \mathcal{Z}$ known
Interpretable BC-IL (Hüyük et al., 2021)	✓	✓	✓	$\mathcal{S}, \mathcal{A}, \mathcal{Z}$ known, low-dim.
DM-IL (Ho & Ermon, 2016)	✗	✗	✗	$\mathcal{S} = \mathcal{Z}$
Interpretable DM-IL (Li et al., 2017)	✗	✗	✓	$\mathcal{S} = \mathcal{Z}$
MB-IL (Englert et al., 2013)	✓	✗	✗	$\mathcal{S} = \mathcal{Z}$; simple dynamics
IRL (Ng & Russell, 2000)	✗	✗	✗	$\mathcal{S} = \mathcal{Z}$; π_E optimal wrt. $\mathcal{R}$
PO-IRL (Choi & Kim, 2011)	✗	✓	✗	π_E optimal wrt. $\mathcal{R}$
Offline PO-IRL (Makino & Takeuchi, 2012)	✓	✓	✗	π_E optimal wrt. $\mathcal{R}$
Interpretable RL (Silva et al., 2020)	✗	✗	✓	$\mathcal{S} = \mathcal{Z}$; $\mathcal{R}$ known
PoETree **(Us)**	✓	✓	✓	$\mathcal{A}, \mathcal{Z}$ known

1996; Piot et al., 2014); and distribution-matching (DM) methods, where state-action distributions between demonstrator and learned policies are matched through adversarial learning (Ho & Ermon, 2016; Jeon et al., 2018; Kostrikov et al., 2020). *Apprenticeship learning* (AL) aims to match or surpass expert performance on a given task, a closely related but distinct problem. The state-of-the-art solution is inverse reinforcement learning (IRL), which recovers R from demonstration trajectories (Ng & Russell, 2000; Abbeel & Ng, 2004; Ziebart et al., 2008); followed by a forward reinforcement learning algorithm to obtain the policy.

Towards interpretability for understanding decisions Imitation and apprenticeship learning solutions have been developed to tackle the challenges of fully-offline learning in partially-observable decision environments, as required by the clinical context: Sun et al. (2017) propose to learn belief states through recurrent neural networks for IL, and IRL was adapted for partially-observable environments (Choi & Kim, 2011) and even to learn dynamics models offline (Makino & Takeuchi, 2012). However, while black-box algorithms such as neural networks typically provide best results, a transparent description of demonstrated behaviour requires parametrising the learned policy to be informative and understandable by a human. This often involves sacrificing action-matching performance for interpretability (Lage et al., 2018), or introduces assumptions which limit the applicability of the proposed method, such as Markovianity or low-dimensional state spaces (Hüyük et al., 2021). Hüyük et al. (2021), for example, share our goal and parametrise action choice through decision boundaries within the agent's belief space. In addition to leaving the mapping from observations to decisions obscure, this imposes restrictive assumptions on belief states and dynamics – which limit predictive performance, scalability and fidelity to the human thought process (Biasutti et al., 2012).

One approach to achieve interpretability in decision models is to implement *interpretable reward functions* for IRL, providing insight into demonstrating agents' inner objectives. Behaviour has thus been explained as preferences over counterfactual outcomes (Bica et al., 2021), or over information value in a time-pressured context (Jarrett & van der Schaar, 2020). Still, the need for black-box RL algorithms to extract a policy obscures how observations affect action choice. Instead, the agent's policy function can be directly *parametrised as an interpretable structure*. RL policies have been represented through a high-level programming syntax (Verma et al., 2018), or through explanations in terms of intended outcomes (Yau et al., 2020).

Moving from a focus on policies to interpretability in time-series generally, the state-of-the-art is to highlight the relative importance of features in contributing to changes in model predictions. This can be achieved with attention mechanisms as in Choi et al. (2016), or with Shapley explanations (Lundberg et al., 2017). Still, these methods remain tools to gain approximative insight into complex models, rather than a transparent description of their inner workings. In addition, patient observations and medical actions are all treated as independent variables – a significant assumption for policy learning, making models susceptible to confounding in observational data. In contrast, our work aims to distill decision-making pathways from observations to actions in a clear format, comparable to the human thought process.

Adaptive Decision Trees for Time-Series Related soft and grown tree architectures have been proposed for image-based tasks and visualisation (Frosst & Hinton, 2017; Tanno et al., 2019) – little research has been conducted on developing probabilistic decision trees for the heterogeneous, sparse nature of tabular data. Our work is further contrasted with established tree models in Appendix

B. More complex tree structures with improved modelling power have also been reported, such as 'transformer' elements which project input variables to more informative features in between decision nodes (Tanno et al., 2019) in a form of representation learning. Ding et al. (2021) also carry out representation learning through a cascade of trees, sequentially using their outputs as input to subsequent ones, up to a final action-selection structure. In all cases, however, models become difficult to understand, as decisions are based on unintuitive variables: in contrast, we build thresholds over native observation variables, meaningful to human experts.

5 Illustrative Examples

5.1 Experimental Setup

Decision-Making Environments A synthetic dataset of 1000 patient trajectories over 9 timesteps, denoted **SYNTH**, was first simulated to assess effectiveness in recovering ground-truth policies π_E, inaccessible in real data. In a PO-MDP with binary disease state space, observations $\{z_+,\ z_-\}$ encode diagnostic tests and also include 10 dimensions disregarded by the agent (e.g. irrelevant clinical measurements); and actions $\{a_+,\ a_-\}$ correspond to treatment and lack thereof. Assuming treatment efficacy requires exposure beyond the end of symptoms (Entsuah et al., 1996), the expert policy is to treat patients who tested positive over the last three timesteps: $a_t = \pi_E(h_t, z_t) = a_+$ if $z_+ \in \{z_{t-2}, z_{t-1}, z_t\}$; a_- else.

Our first medical dataset consists of bi-annual data for 1,625 patients collected for the Alzheimer's Disease Neuroimaging Initiative (**ADNI**) database. We consider the task of predicting whether an MRI scan is ordered (action a_t) at each visit: neuroimaging is useful to diagnose Alzheimer's disease and causes of cognitive decline (NICE, 2018). Patient observations consist of the Clinical Dementia Rating (CDR-SB) on a severity scale (normal; questionable impairment; mild/severe dementia) following O'Bryant et al. (2008); and the MRI outcome of the previous visit, categorised into four possibilities (no MRI scan; and below-average, average and above-average hippocampal volume V_h). We also consider a dataset of 4,222 ICU patients over 6 timesteps extracted from the third Medical Information Mart for Intensive Care (**MIMIC-III**) database (Johnson et al., 2016) and predict antibiotic prescription based on 8 time-dependent observations – temperature, white blood cell count, heart rate, hematocrit, hemoglobin, blood pressure, creatinine and potassium, following Bica et al. (2021).

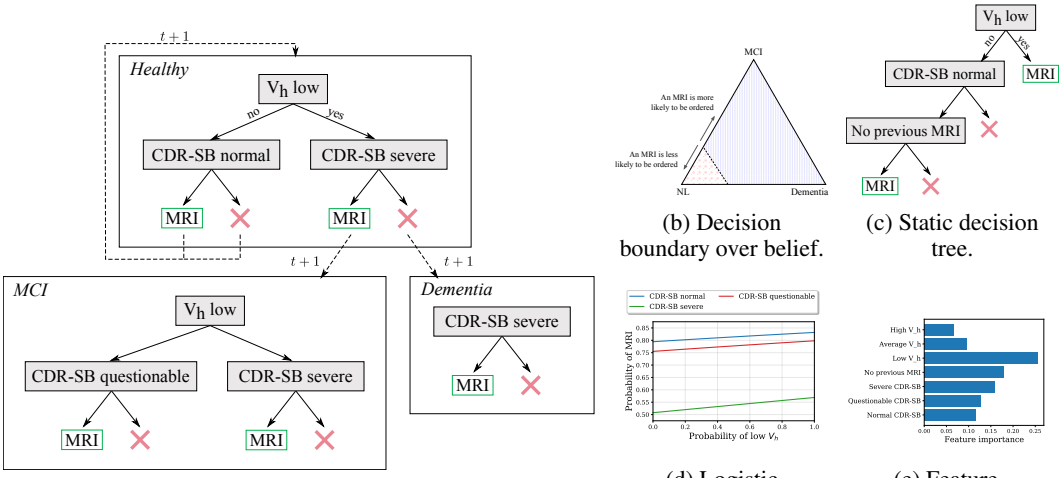
Success Metrics Reflecting our unconventional priority for *policy interpretability* over imitation performance, we provide illustrative examples evidencing greater insight into observed behaviour through decision tree representations. For a more objective assessment, we also surveyed XXX practising UK clinicians of different seniority level for feedback on policy interpretability and sensibility. *Action-matching performance* was also evaluated through metrics accounting for class imbalance: area under the receiver-operating-characteristic curve (AUROC), under the precision-recall curve (AUPRC), and Brier calibration (better, i.e. lower, scores for uncertain predictions when incorrect).

Benchmark Algorithms Different behavioural cloning models, mapping observations or history to actions, are implemented for benchmarking: (i) a decision tree with no handling of patient history (**Interpretable BC-IL**); (ii) a partially-observable BC algorithm (**PO-BC-IL**) extracting h_t from observations $z_{1:t}$ through an RNN (Sun et al., 2017); (iii) and an **Interpretable PO-BC-IL** model (Hüyük et al., 2021) learning PO-MDP dynamics through an Input-Output Hidden Markov Model (IOHMM) (Bengio & Frasconi, 1995) and parametrising the policy by dividing the belief space into regions where different actions are more probable. Additional benchmarks include (iv) model-based imitation learning **PO-MB-IL** (Englert et al., 2013), adapted for partially-observability with a learned IOHMM environment model; (v) Bayesian Inverse Reinforcement Learning for PO-MDPs (**PO-IRL**), based on Jarrett & van der Schaar (2020); and (vi) a **Offline PO-IRL** version of this algorithm as in Makino & Takeuchi (2012). Implementation details are provided in Appendix E.

5.2 Interpretability

We propose several clinical examples and insights from the ADNI dataset to highlight the usefulness of our policy representation, in comparison to the state-of-the-art in interpretable policy learning.

Explaining patient trajectories Our first example studies decision-making within typical ADNI trajectories. Let **Patient A** be healthy with normal CDR score and average V_h on initial scan. Let **Patient B** initially have mild cognitive impairment (MCI), progressing towards dementia. Their



(a) Adaptive decision tree policy

(b) Decision boundary over belief.

(c) Static decision tree.

(d) Logistic Regression.

(e) Feature importance.

Figure 2: **Comparison of interpretable policy learning methods on ADNI.** Red crosses stand for no action. Figures (b) and (e) follow methods from Hüyük et al. (2021) and Lundberg et al. (2017).

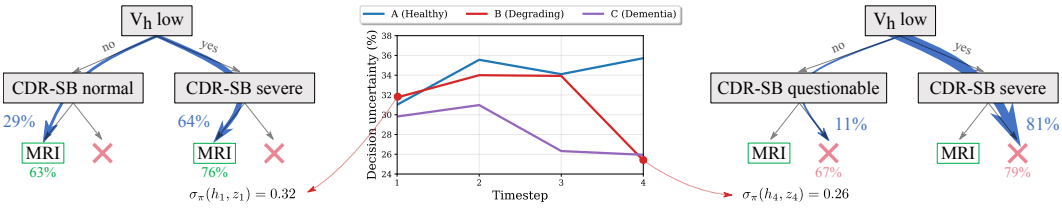


Figure 3: **Policy decision uncertainty** σ_π for ADNI[1]. Maximum probability paths are highlighted in blue, with action probabilities in green/red.

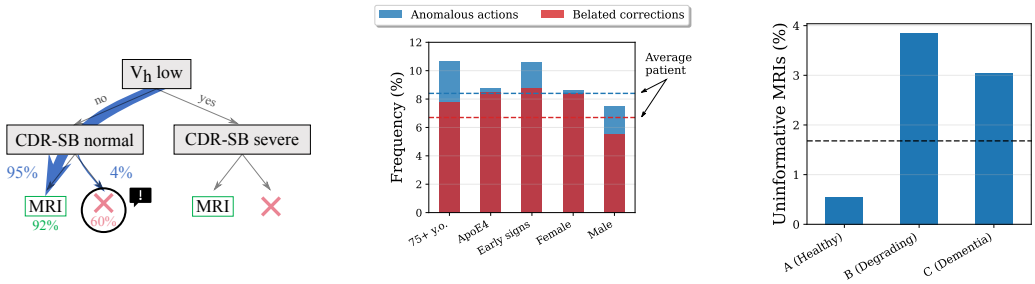
CDR degrades from questionable to severe at the last timestep and a below-average V_h is measured at each visit. **Patient C** is diagnosed with dementia at the first visit, with severe dementia CDR and low V_h: no further scans are ordered as they would be uninformative (NICE, 2018). Figure 2a illustrates our learned tree policy, both varying with and determining patient history at each timestep. Patient A is not ordered any further scan, as they never show low V_h or abnormal CDR, but may be ordered one if they develop cognitive impairment. For Patient B with low V_h but a CDR score not yet severe, another MRI is ordered to monitor disease progression. At the following visit, patient history conditions the physician to be more cautious of scan results: even if V_h is found to be acceptable, another scan may be ordered if the CDR changes. Finally, Patient C's observations already give a certain diagnosis of dementia: no scan is necessary. At following timesteps, unless the patient unexpectedly recovers, their history suggests to *not* carry out any further scans. We must highlight the similarity between our decision tree policy and *published guidelines* for Alzheimer diagnosis in Biasutti et al. (2012), reproduced in Appendix F.

In contrast, the black-box history extraction procedure required by most benchmarks for partially-observability (PO-BC-IL, PO-IRL, Offline PO-IRL, PO-MB-IL) is difficult to interpret, with no clear decision-making process. The policy learned by Interp. PO-BC-IL (Hüyük et al., 2021) in Figure 2b, represented through decision boundaries over a subjective disease belief space, does provide more insight. This model correctly suggests few or frequent scans for healthy Patient A and degrading Patient B respectively, but cannot account for scan reduction in diagnosed Patient C as this requires discontinuous action regions over the belief state. This policy representation thus cannot capture the effect of different observations and histories on treatment strategies, independently of beliefs in patient disease state.

Decision-making uncertainty Our behaviour model also inherently captures the useful notion of decision uncertainty thanks to its probabilistic nature, as:

$$\sigma_\pi(h_t, z_t) = 1 - \sum_{l \in LeafNodes} P^l(h_t, z_t) \cdot Q_a^l \tag{5}$$

[1]Can be recovered approximately by inputting values in equation 5: $1-(0.29\times0.63+0.64\times0.76) = 0.33$. Exact value obtained with all leaves.



(a) Anomalous behaviour.

(b) Anomalous behaviour per cohort.

(c) Uninformative actions.

Figure 4: **Anomalous and low-value actions in various patient cohorts**: patients over 75 years old, patients carrying the $\epsilon 4$ allele of the apolipoprotein E gene (ApoE4), patients with signs of MCI or dementia at their first visit, female and male patients.

where, for each leaf node l, $P^l(h_t, z_t)$ and Q^l_a are the path probability and output probability of the chosen action $a = \arg\max_a \pi(a_t = a|h_t, z_t)$. Figure 3 highlights how action uncertainty varies over time for each typical patient. High uncertainty for Patient A can be attributed to variability in healthy patient care in the demonstrations dataset, as scans may provide information about their state, but must be balanced with cost and time considerations (NICE, 2018). For Patient C with a dementia diagnosis, further investigations are not required (Biasutti et al., 2012), reflected in uncertainty reduction. As a result, for Patient B with degrading symptoms, scans are initially ordered with high uncertainty to monitor progression, but as symptoms worsen and the patient is diagnosed, decision uncertainty reduces. This example illustrates the value of uncertainty estimation afforded by our framework, modulating probability values within tree paths and leaf outputs.

Anomalous behaviour detection Uncovered policies are also valuable to flag actions incompatible with the decision-making behaviour, defined as a lack of scan when this was the recommended action with 90% certainty as in Figure 4a – making up 8.4% of ADNI. This highlights flaws in demonstrating behaviour, as the learned policy has a low Brier score in Table 2, a sign of well-calibrated probability estimates, and is here confident in its prediction. Anomalous behaviour *followed* by an MRI at the next visit, 'correcting' off-policy actions and resulting in belated diagnoses, can also be detected. Results suggest 6.7% of patients are thus investigated late – comparable to the belated diagnoses found in 6.5% of patients by Hüyük et al. (2021). As shown in Figure 4b, high-risk cohorts due to age, female gender or ApoE4 allele (Launer et al., 1999; Lindsay et al., 2002) are more often subject to late actions. Most anomalous actions on patients with ApoE4 or female patients end up corrected, in contrast to older patients or ones with early symptoms – imaging is less paramount for the latter, as they may be diagnosed clinically (NICE, 2018).

Action value quantification through counterfactual outcomes Our joint expected outcome prediction and policy learning model allows action value to be quantified (Yau et al., 2020; Bica et al., 2021). Low-value actions are defined as a scan being ordered and showing similar observations to the previous timestep, when this could already be foreseen from the expected outcome (less than average variation minus one standard deviation). As shown in Figure 4c, ill patients such as B and C are monitored more closely than healthy ones – particularly B with uncertain diagnosis – although not always by necessity. This analysis agrees again with previous work which found a high rate of uninformative actions for patients with early signs of disease (Hüyük et al., 2021). Overall, our framework makes the assessment of factual and counterfactual evolution intuitive to external users as it is integrated in the modelling process by design.

5.3 Policy Fidelity

Action-matching performance As shown in Table 2, our approach outperforms or comes second-best *on all success metrics* compared to benchmark algorithms. This confirms the expressive modelling power of our recurrent tree policies over decision boundaries (Hüyük et al., 2021) and even over neural-network-based methods (Sun et al., 2017), specifically optimised for action-matching. Mean relative outcome error was measured as $13 \pm 4\%$ and $26 \pm 9\%$ on ADNI and MIMIC respectively, giving reasonable and insightful values. As discussed previously, our survey of end-users also suggests our model is the most understandable, with more details in Appendix G.

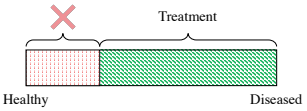
Table 2: **Comparison of the performance of policy learning algorithms[2] on medical datasets.** Interpretability scores out of ten were obtained through our clinician survey. Lower is better for Brier calibration. Standard errors for MIMIC and SYNTH were ≤ 0.04.

Task	ADNI MRI scans				MIMIC antibiotics		SYNTH	
Algorithm	Interpretability	AUROC	AUPRC	Brier	AUC-ROC	Brier	AUROC	Brier
Interp. BC-IL	9.3 ± 0.3	0.53 ± 0.01	0.72 ± 0.01	0.25 ± 0.01	0.50	0.23	–	–
PO-BC-IL [49]	0.3 ± 0.3	0.59 ± 0.04	0.80 ± 0.08	0.18 ± 0.05	0.67	0.16	0.98	0.02
Interp. PO-BC-IL [22]	7.3 ± 0.5	0.44 ± 0.04	0.75 ± 0.09	0.19 ± 0.07	0.65	0.21	0.84	0.12
PO-MB-IL [14]	–	0.54 ± 0.08	0.81 ± 0.07	0.19 ± 0.03	–	–	–	–
PO-IRL [24]	–	0.50 ± 0.08	0.82 ± 0.04	0.23 ± 0.01	–	–	–	–
Offline PO-IRL [38]	–	0.54 ± 0.06	0.83 ± 0.04	0.23 ± 0.01	–	–	–	–
PoETree	$\mathbf{8.3 \pm 0.3}$	$\mathbf{0.62 \pm 0.01}$	$\mathbf{0.82 \pm 0.01}$	$\mathbf{0.18 \pm 0.01}$	**0.68**	**0.19**	**0.99**	**0.01**

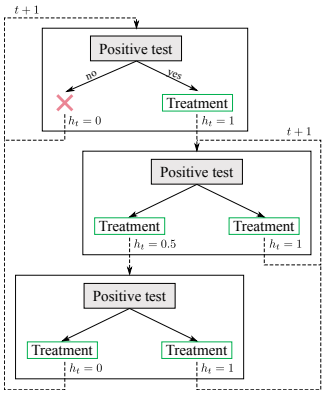
Policy identifiability Figure 5 illustrates policies learned from our proof-of-concept SYNTH dataset with PoETree and with our closest related work (Hüyük et al., 2021). Despite the apparent simplicity of the decision-making problem, the latter failed to identify the outlined treatment strategy, as evidenced by its inferior performance in Table 2. This illustrates greater flexibility of our tree representation in disambiguating decision-making factors which simply cannot be captured by related work.

Inductive bias Having demonstrated that our policy learning framework allows us to *recover* domain expertise, we investigated the effect of incorporating *prior* policy knowledge on learning performance. In practice, our model's action prediction structure is initialised based on a published guideline (Biasutti et al., 2012) shown in Appendix F and optimised: training time was reduced by a factor of XX and action-matching AUPRC was improved to 84.0 $\pm$ 0.5 %. This 'warm-start' is a promising strategy to learn from physician observation while building on established clinical practice standards.

Accuracy-interpretability trade-off: complexity analysis Incrementally grown trees outperform fixed structures by eliminating unnecessary partitions, with $\sim$40% fewer parameters for the same depth, greater readability (Lage et al., 2018) and reasonable training time. Grown structures adapt their depth and complexity to avoid overfitting while capturing the modelling task, in a form of Occam's razor – jointly optimising policy accuracy and interpretability. In particular, policies adapt to sample complexity in training. Detailed experimental results are provided in Appendix D.



(a) Decision-boundary policy (Hüyük et al., 2021).



(b) Adaptive decision tree policy.

Figure 5: **Recovered policies for SYNTH.**

6 Conclusion

Established paradigms to replicate demonstrated behaviour fall short in explaining learned decision policies, whereas the state-of-the-art in interpretable policy learning was criticised by end-users for restrictive modelling assumptions (Hüyük et al., 2021) – thus not fully capturing diagnostic strategies. Policy Extraction with Decision Trees (PoETree) overcomes these challenges, outputting most likely actions and uncertainty estimates, given patients' observations and medical history. Our multi-output tree model handles history and predicts expected outcomes through recurrence – justifying decisions by comparing counterfactual evolution. Our method outperforms the state-of-the-art in interpretability, policy fidelity and action-matching, thus providing an understanding demonstrated behaviour without compromising accuracy. In particular, we can formalise and quantify concepts of practice variation with time, anomalous and belated behaviour, and low-value actions; and validate clinical insights through a survey of intended users.

On a bigger picture, our novel interpretable framework for time-series modelling shows promise for alternative tasks relevant to clinical decision support, such as patient trajectory prediction or human-in-the-loop learning – which we intend to explore in further work.

[2]References: [49] Sun et al. (2017); [22] Hüyük et al. (2021); [14] Englert et al. (2013); [24] Jarrett & van der Schaar (2020); [38] Makino & Takeuchi (2012).

ACKNOWLEDGMENTS

We thank all NHS clinicians who participated in our survey and the Alzheimer's Disease Neuroimaging Initiative. MIMIC-III

REFERENCES

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. *International Conference on Machine Learning (ICML)*, 2004.

Michael Bain and Claude Sammut. A framework for behavioural cloning. *MACHINE INTELLIGENCE 15*, pp. 103–129, 1996. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.25.1759.

AH Beck, AR Sangoi, S Leung, RJ Marinelli, TO Nielsen, MJ van de Vijver, RB West, M van de Rijn, and D Koller. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science translational medicine*, 3, 11 2011. ISSN 1946-6242. doi: 10.1126/SCITRANSLMED.3002564. URL https://pubmed.ncbi.nlm.nih.gov/22072638/.

Yoshua Bengio and Paolo Frasconi. An input output hmm architecture. *Adv. Neural Inf. Process. Syst.*, 7:427–434, 1995.

Maria Biasutti, Natacha Dufour, Clotilde Ferroud, William Dab, and Laura Temime. Cost-effectiveness of magnetic resonance imaging with a new contrast agent for the early diagnosis of alzheimer's disease. *PLOS ONE*, 7:e35559, 4 2012. ISSN 1932-6203. doi: 10.1371/JOURNAL.PONE.0035559. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0035559.

Ioana Bica, Daniel Jarrett, Alihan Hüyük, and Mihaela van der Schaar. Learning "what-if" explanations for sequential decision-making. *Proc. 9th International Conference on Learning Representations (ICLR 2021)*, 2021. URL http://arxiv.org/abs/2007.13531.

Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees.* Wadsworth & Brooks/Cole, 1984.

J Burch, S Hinde, S Palmer, F Beyer, J Minton, A Marson, U Wieshmann, N Woolacott, and M Soares. The clinical effectiveness and cost-effectiveness of technologies used to visualise the seizure focus in people with refractory epilepsy being considered for surgery: a systematic review and decision-analytical model. *Health Technology Assessment*, 16:1–163, 9 2012. ISSN ISSN: 2046-4924. doi: 10.3310/HTA16340. URL www.hta.ac.uk.

Randall S. Burd and Frank A. Sonnenberg. Decision analysis: A basic overview for the pediatric surgeon. *Seminars in Pediatric Surgery*, 11:46–54, 2 2002. ISSN 1055-8586. doi: 10.1053/SPSU.2002.29370.

Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Avances in Neural Information Processing Systems (NIPS)*, 29, 8 2016. URL http://arxiv.org/abs/1608.05745.

Jaedeug Choi and Kee-Eung Kim. Inverse reinforcement learning in partially observable environments. *Journal of Machine Learning Research*, 12:691–730, 2011. ISSN 1533-7928. URL http://jmlr.org/papers/v12/choi11a.html.

Roger Chou, Amir Qaseem, Vincenza Snow, Donald Casey, Thomas J. Cross, Paul Shekelle, and Douglas K. Owens. Diagnosis and treatment of low back pain: A joint clinical practice guideline from the american college of physicians and the american pain society. *Annals of Internal Medicine*, 147:478–491, 10 2007. doi: 10.7326/0003-4819-147-7-200710020-00006.

Zihan Ding, Pablo Hernandez-Leal, Gavin Weiguang Ding, Changjian Li, and Ruitong Huang. Cdt: Cascading decision trees for explainable reinforcement learning. *arXiv preprint arXiv:2011.07553*, 2021. URL https://github.com/openai/gym/blob/master/gym/envs/box2d/lunar.

Peter Englert, Alexandros Paraschos, Marc Peter Deisenroth, and Jan Peters. Probabilistic model-based imitation learning. *Adaptive Behavior*, 21:388–403, 10 2013. ISSN 10597123. doi: 10.1177/1059712313491614. URL https://journals.sagepub.com/doi/full/10.1177/1059712313491614.

A. R. Entsuah, R. L. Rudolph, D. Hackett, and S. Miska. Efficacy of venlafaxine and placebo during long-term treatment of depression: A pooled analysis of relapse rates. *International Clinical Psychopharmacology*, 11:137–145, 1996. doi: 10.1097/00004850-199611020-00008. URL /record/1996-05005-007.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature 2017 542:7639*, 542:115–118, 1 2017. ISSN 1476-4687. doi: 10.1038/nature21056. URL https://www.nature.com/articles/nature21056.

Wael Farag. Safe-driving cloning by deep learning for autonomous cars. *International Journal of Advanced Mechatronic Systems*, 7:390–397, 2017. doi: 10.1504/IJAMECHS.2017.099318.

Nicholas Frosst and Geoffrey Hinton. Distilling a neural network into a soft decision tree. *CEX workshop at AI*IA 2017 conference*, 11 2017. URL http://arxiv.org/abs/1711.09784.

Matthew Golub, Steven Chase, and Byron Yu. Learning an internal dynamics model from control demonstration. pp. 606–614, 2 2013. ISSN 1938-7228. URL http://proceedings.mlr.press/v28/golub13.html.

V Gulshan, L Peng, M Coram, MC Stumpe, D Wu, A Narayanaswamy, S Venugopalan, K Widner, T Madams, J Cuadros, R Kim, R Raman, PC Nelson, JL Mega, and DR Webster. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316:2402–2410, 12 2016. ISSN 1538-3598. doi: 10.1001/JAMA.2016.17216. URL https://pubmed.ncbi.nlm.nih.gov/27898976/.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *30th Conference on Neural Information Processing Systems*, 2016.

Alihan Hüyük, Daniel Jarrett, Cem Tekin, and Mihaela Van Der Schaar. Explaining by imitating: Understanding decisions by interpretable policy learning. *ICLR*, 2021.

Ozan Irsoy, Olcay Taner Yıldız, and Ethem Alpaydın. Soft decision trees. *21st International Conference on Pattern Recognition (ICPR)*, 2012.

Daniel Jarrett and Mihaela van der Schaar. Inverse active sensing: Modeling and understanding timely decision-making. *Proceedings of the 37 th International Conference on Machine Learning*, 2020.

Daniel Jarrett, Alihan Hüyük, and Mihaela Van Der Schaar. Inverse decision modeling: Learning interpretable representations of behavior. *Proceedings of the 38th International Conference on Machine Learning*, 139:4755–4771, 7 2021. ISSN 2640-3498. URL http://proceedings.mlr.press/v139/jarrett21a.html.

Wonseok Jeon, Seokin Seo, and Kee-Eung Kim. A bayesian approach to generative adversarial imitation learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li Wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:1–9, 5 2016. ISSN 20524463. doi: 10.1038/sdata.2016.35. URL www.nature.com/sdata/.

Yohannes Kassahun, Bingbin Yu, Abraham Temesgen Tibebu, Danail Stoyanov, Stamatia Giannarou, Jan Hendrik Metzen, and Emmanuel Vander Poorten. Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions. *International Journal of Computer Assisted Radiology and Surgery 2015 11:4*, 11:553–568, 10 2015. ISSN 1861-6429. doi: 10.1007/S11548-015-1305-Z. URL https://link.springer.com/article/10.1007/s11548-015-1305-z.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 12 2015. URL https://arxiv.org/abs/1412.6980v9.

Ilya Kostrikov, Ofir Nachum, and Jonathan Tompson. Imitation learning via off-policy distribution matching. *ICLR*, 12 2020. URL http://arxiv.org/abs/1912.05032.

Isaac Lage, Andrew Slavin Ross, Been Kim Google Brain, Samuel J Gershman, and Finale Doshi-Velez. Human-in-the-loop interpretability prior. *32nd Conference on Neural Information Processing Systems (NIPS 2018)*, 2018.

LJ Launer, K Andersen, ME Dewey, L Letenneur, A Ott, LA Amaducci, C Brayne, JR Copeland, JF Dartigues, P Kragh-Sorensen, Lobo A, Martinez-Lage JM, T Stijnen, and A Hofman. Rates and risk factors for dementia and alzheimer's disease: results from eurodem pooled analyses. eurodem incidence research group and work groups. european studies of dementia. *Neurology*, 52:78–84, 1 1999. ISSN 0028-3878. doi: 10.1212/WNL.52.1.78. URL https://pubmed.ncbi.nlm.nih.gov/9921852/.

M.-C Laï, M Brian, and M.-F Mamzer. Perceptions of artificial intelligence in healthcare: findings from a qualitative survey study among actors in france. *J Transl Med*, 18:14, 2020. doi: 10.1186/s12967-019-02204-y. URL https://doi.org/10.1186/s12967-019-02204-y.

A Li, S Jin, L Zhang, and Y Jia. A sequential decision-theoretic model for medical diagnostic system. *Technology and health care : official journal of the European Society for Engineering and Medicine*, 23 Suppl 1:S37–S42, 5 2015. ISSN 1878-7401. doi: 10.3233/THC-150926. URL https://pubmed.ncbi.nlm.nih.gov/26410326/.

Yunzhu Li, Jiaming Song, and Stefano Ermon. Infogail: Interpretable imitation learning from visual demonstrations. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 3815–3825, 2017. URL https://github.com/ermongroup/InfoGAIL.

J Lindsay, D Laurin, R Verreault, R Hébert, B Helliwell, GB Hill, and I McDowell. Risk factors for alzheimer's disease: a prospective analysis from the canadian study of health and aging. *American journal of epidemiology*, 156:445–453, 9 2002. ISSN 0002-9262. doi: 10.1093/AJE/KWF074. URL https://pubmed.ncbi.nlm.nih.gov/12196314/.

Scott M Lundberg, Paul G Allen, and Su-In Lee. A unified approach to interpreting model predictions. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017. URL https://github.com/slundberg/shap.

Takaki Makino and Johane Takeuchi. Apprenticeship learning for model parameters of partially observable environments. *Proceedings of the 29 th International Conference on Machine Learning*, pp. 891–898, 2012.

R. G. Masterton, A. Galloway, G. French, M. Street, J. Armstrong, E. Brown, J. Cleverley, P. Dilworth, C. Fry, A. D. Gascoigne, Alan Knox, Dilip Nathwani, Robert Spencer, and Mark Wilcox. Guidelines for the management of hospital-acquired pneumonia in the uk: Report of the working party on hospital-acquired pneumonia of the british society for antimicrobial chemotherapy. *Journal of Antimicrobial Chemotherapy*, 62:5–34, 7 2008. ISSN 0305-7453. doi: 10.1093/JAC/DKN162. URL https://academic.oup.com/jac/article/62/1/5/844812.

Ann M. McCreery and Edmond Truelove. Decision making in dentistry. part i: A historical and methodological overview. *The Journal of Prosthetic Dentistry*, 65:447–451, 3 1991. ISSN 0022-3913. doi: 10.1016/0022-3913(91)90241-N.

J. B. McKinlay, C. L. Link, K. M. Freund, L. D. Marceau, A. B. O'Donnell, and K. L. Lutfey. Sources of variation in physician adherence with clinical guidelines: Results from a factorial experiment. *Journal of General Internal Medicine*, 22:289, 3 2007. doi: 10.1007/S11606-006-0075-2. URL /pmc/articles/PMC1824760//pmc/articles/PMC1824760/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC1824760/.

Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. *Proc. 17th International Conf. on Machine Learning*, pp. 663–670, 2000.

NICE. Dementia: Assessment, management and support for people living with dementia and their carers. *National Institute for Health and Care Excellence (NICE) Guideline No. 97*, 2018. URL http://www.nice.org.uk/guidance/ng97.

SE O'Bryant, SC Waring, CM Cullum, J Hall, L Lacritz, PJ Massman, PJ Lupo, JS Reisch, and R Doody. Staging dementia using clinical dementia rating scale sum of boxes scores: a texas alzheimer's research consortium study. *Archives of neurology*, 65:1091–1095, 8 2008. ISSN 1538-3687. doi: 10.1001/ARCHNEUR.65.8.1091. URL https://pubmed.ncbi.nlm.nih.gov/18695059/.

Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, and Jan Peters. An algorithmic perspective on imitation learning. *Foundations and Trends ® in Robotics*, 7: 1–179, 2018. doi: 10.1561/2300000053.

Bilal Piot, Matthieu Geist, and Olivier Pietquin. Boosted and reward-regularized classification for apprenticeship learning. *Proceedings of the 13th Inter-national Conference on Autonomous Agents and MultiagentSystems (AAMAS 2014)*, 2014. URL www.ifaamas.org.

Amir Qaseem. Management of stable ischemic heart disease: Summary of a clinical practice guideline from the american college of physicians/american college of cardiology foundation/american heart association/american association for thoracic surgery/preventive cardiovascular nurses association/society of thoracic surgeons. *Annals of Internal Medicine*, 157: 735, 11 2012. ISSN 0003-4819. doi: 10.7326/0003-4819-157-10-201211200-00011. URL https://pubmed.ncbi.nlm.nih.gov/23165665/.

Andrew Silva, Taylor Killian, Ivan Rodriguez Jimenez, Sung-Hyun Son, and Matthew Gombolay. Optimization methods for interpretable differentiable decision trees in reinforcement learning. *Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Wen Sun, Arun Venkatraman, Geoffrey J Gordon, Byron Boots, and J Andrew Bagnell. Deeply aggrevated: Differentiable imitation learning for sequential prediction. *Proceedings of the 34th International Conference on Machine Learning*, pp. 3309–3318, 7 2017. ISSN 2640-3498. URL http://proceedings.mlr.press/v70/sun17d.html.

Ryutaro Tanno, Kai Arulkumaran, Daniel C Alexander, Antonio Criminisi, and Aditya Nori. Adaptive neural trees. *Proceedings of the 36th International Conference on Machine Learning*, pp. 6166–6175, 5 2019. ISSN 2640-3498. URL http://proceedings.mlr.press/v97/tanno19a.html.

The Royal Society. Machine learning: The power and promise of computers that learn by example. *The Royal Society*, 2017.

Gilmer Valdes, Charles B. Simone, Josephine Chen, Alexander Lin, Sue S. Yom, Adam J. Pattison, Colin M. Carpenter, and Timothy D. Solberg. Clinical decision support of radiotherapy treatment planning: A data-driven machine learning strategy for patient-specific dosimetric decision making. *Radiotherapy and Oncology*, 125:392–397, 12 2017. ISSN 0167-8140. doi: 10.1016/J.RADONC.2017.10.014.

Abhinav Verma, Vijayaraghavan Murali, Rishabh Singh, Pushmeet Kohli, and Swarat Chaudhuri. Programmatically interpretable reinforcement learning. *35th International Conference on Machine Learning, ICML 2018*, 11:8024–8033, 4 2018. URL https://arxiv.org/abs/1804.02477v3.

Gert P. Westert, Stef Groenewoud, John E. Wennberg, Catherine Gerard, Phil Dasilva, Femke Atsma, and David C. Goodman. Medical practice variation: Public reporting a first necessary step to spark change. *International Journal for Quality in Health Care*, 30:731–735, 11 2018. ISSN 14643677. doi: 10.1093/intqhc/mzy092. URL /pmc/articles/PMC6307331//pmc/articles/PMC6307331/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6307331/.

Herman Yau, Chris Russell, and Simon Hadfield. What did you think would happen? explaining agent behaviour through intended outcomes. *Advances in Neural Information Processing Systems*, 33:18375–18386, 2020. URL https://github.com/hmhyau/rl-intention.

Brian D Ziebart, Andrew Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, 2008. URL www.aaai.org.

Ariel Zylberberg, Jeannette AM Lorteije, Brian G Ouellette, Chris I De Zeeuw, Mariano Sigman, and Pieter Roelfsema. Serial, parallel and hierarchical decision making in primates. *eLife*, 6, 6 2017. doi: 10.7554/ELIFE.17331. URL /pmc/articles/PMC5484613//pmc/articles/PMC5484613/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5484613/.

A Additional Related Work

A.1 Imitation learning

A.1.1 Behavioural cloning

Behavioural cloning (BC) is the simplest form imitation learning, as it reduces policy learning to supervised learning of a direct mapping from states to actions, assuming a fully-observable setting where $\mathcal{S} = \mathcal{Z}$ (Bain & Sammut, 1996; Piot et al., 2014). For a trajectories dataset $\mathcal{D} = \{(z_{1:T}, a_{1:T})\}$, the objective is to recover the deterministic policy mapping $\pi : \mathcal{Z} \to \mathcal{A}$, such that $\pi(z_t)$ is the action following observation z_t:

$$\pi = \arg\min_{\pi} \mathbb{E}_{z_t, a_t \sim \mathcal{D}} \left[\mathcal{L}\left(\pi(z_t),\ a_t\right) \right] \tag{6}$$

where $\mathcal{L}(\cdot, y)$ is a supervised learning loss function for target value y.

This approach is typically compatible with offline learning, as it only requires access to demonstrated trajectories of system states and actions (z_t, a_t) at each timestep t. Its excellent learning efficiency has motivated its use and deployment, including in safety-critical applications such as autonomous driving (Farag, 2017).

With no explicit handling of system dynamics, however, this approach is highly susceptible to distributional shift (Ho & Ermon, 2016): at test time, the agent might enter states unseen in training trajectories and thus drift away from its training domain and perform poorly. Likewise, BC is not designed for non-Markovian or partially-observable environments, where the latest observation z_t is insufficient to determine action choice. Representation learning of PO-MDP belief states through recurrent neural networks (Sun et al., 2017) can help tackle these issues at the cost of interpretability, as it introduces black-box elements in the model.

A.1.2 State-action distribution matching

More recently, an alternative approach to dealing with covariate shift was proposed, based on matching state-action distributions between the demonstrator and learned policies. Defining occupancy measures $\mu_\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for each policy π as $\mu_\pi(s, a) = \pi(a|s) \sum_{t=0}^{\infty} \gamma^t P(s_t = s|\pi)$, this can be achieved by minimising the Jensen-Shannon divergence in occupancy measures for each policy, while maximising the entropy of the learned one $H(\pi)$ through generative adversarial training (Ho & Ermon, 2016; Jeon et al., 2018; Kostrikov et al., 2020):

$$\arg\min_{\pi} D_{\text{JS}}\left(\mu_\pi \| \mu_{\pi_E}\right) \qquad \text{and} \qquad \arg\max_{\pi} H(\pi) \tag{7}$$

In contrast to our descriptive goal, this approach therefore focuses on matching feature expectations rather than explicitely recovering the demonstrated policy mapping from observations or belief states to actions, based on the assumption that this behaviour is optimal – and thus aims to maximise performance on a given task. In addition, it suffers from the same issue as BC-IL in providing interpretable representations of belief states in partially-observable settings, despite Li et al. (2017) improving explainability by clustering similar demonstration trajectories.

In all cases, online roll-outs of the learned policy are required to train a dynamics model of the environment, as $P(s_t = s|\pi)$ is needed to estimate occupancy, making this work incompatible

with purely observational clinical data. To overcome this, model-based imitation learning (MB-IL) proposes to train autoregressive exogenous models of environment dynamics (Englert et al., 2013) for offline feature expectation. Encouraging results were obtained for fully-observable robotic applications, but these dynamics remain excessively simple for the healthcare context.

A comprehensive reviews of imitation learning methods can be found in Osa et al. (2018).

A.2 Apprenticeship learning

Apprenticeship learning is another approach to modelling expert behaviour, concerned with matching the performance of the demonstrating expert in terms of some unknown utility measure R. The state-of-the-art solution to this task is inverse reinforcement learning (IRL), which involves recovering this reward function from demonstration trajectories (Ng & Russell, 2000) – identifying the expert's hidden objectives. A forward reinforcement learning (RL) algorithm is subsequently applied to obtain the policy.

The inverse task of uncovering the reward function is fundamentally ill-posed, as the observed behaviour can be optimal under a multitude of reward signals. Additional constraints must thus be specified to obtain a unique solution, for instance by maximising policy entropy (Ziebart et al., 2008; Choi & Kim, 2011) as described above for distribution-matching imitation learning, or by maximising the margin between induced policies in terms of feature expectations (Ng & Russell, 2000; Abbeel & Ng, 2004):

$$\arg\min_{\pi} \mathbb{E}_{s\sim\mu_{\pi_E}}[V^{\pi}(s) - V^{\pi_E}(s)] \tag{8}$$

where $V^{\pi}(s)$ is a parametrisation of the value function for policy π, the expected sum of total discounted rewards following this policy, assuming a start in state s: $V^{\pi}(s) = \mathbb{E}_{\pi}[\sum_{t=0}^{\infty} \gamma^t R(s_t)|s_0 = s]$.

IRL can deal with partially-observable environments (Choi & Kim, 2011) and even learn dynamics models fully offline (Makino & Takeuchi, 2012). Still, in terms of providing insight into the decision-making process, the indirect nature of the IRL $\circ$ RL framework results in uninterpretable policies – despite attempts to parametrise reward functions for explainability (Bica et al., 2021).

Table 3: **Comparison of related work in light of our key policy learning goals.** DM-IL stands for distribution-matching imitation learning. f stands for a representation learning step; b_t is the agent's belief over the true state s_t.

Related work	Optimisation Objective	Offline Learning	Partial Observability	Interpretable Policy
BC-IL (Bain & Sammut, 1996)	$\mathcal{L}(\arg\max_a \pi(a\|z_t), a_t)$	✓	✗	✗
PO-BC-IL (Sun et al., 2017)	$\mathcal{L}(\arg\max_a \pi(a\|b_t), a_t); b_t = f(z_{1:t})$	✓	✓	✗
Interpretable BC-IL (Hüyük et al., 2021)		✓	✓	✓
DM-IL (Ho & Ermon, 2016)	$D_{KL}(\pi(a_t\|z_t)\|\pi_E(a_t\|z_t))$	✗	✗	✗
Interpretable DM-IL (Li et al., 2017)		✗	✗	✓
MB-IL (Englert et al., 2013)		✓	✗	✗
IRL (Ng & Russell, 2000)	$\pi = \arg\min_{\pi} \mathbb{E}_{z_t\sim\mu_{\pi_E}}[V^{\pi}(z_t) - V^{\pi_E}(z_t)]$	✗	✗	✗
PO-IRL (Choi & Kim, 2011)	$\pi = \arg\min_{\pi} \mathbb{E}_{b_t\sim\mu_{\pi_E}}[V^{\pi}(b_t) - V^{\pi_E}(b_t)]; b_t = f(z_{1:t})$	✗	✓	✗
Offline PO-IRL (Makino & Takeuchi, 2012)		✓	✓	✗
Subjective PO-IRL (Golub et al., 2013)		✗	✓	✗

Table 3 compares how related research addresses the key goals outlined for our work. Our policy learning algorithm must be operable offline on observational data; must accommodate a partially-observable environment to allow adaptive behaviour over time; and must result in an interpretable representation of the policy. In comparison to Hüyük et al. (2021) which addresses all three goals, we propose to implement a policy learning algorithm with fewer assumptions and restrictions on the belief state and dynamics model structure, as well as a more interpretable and expressive policy parametrisation.

B Soft Decision Tree Optimisation

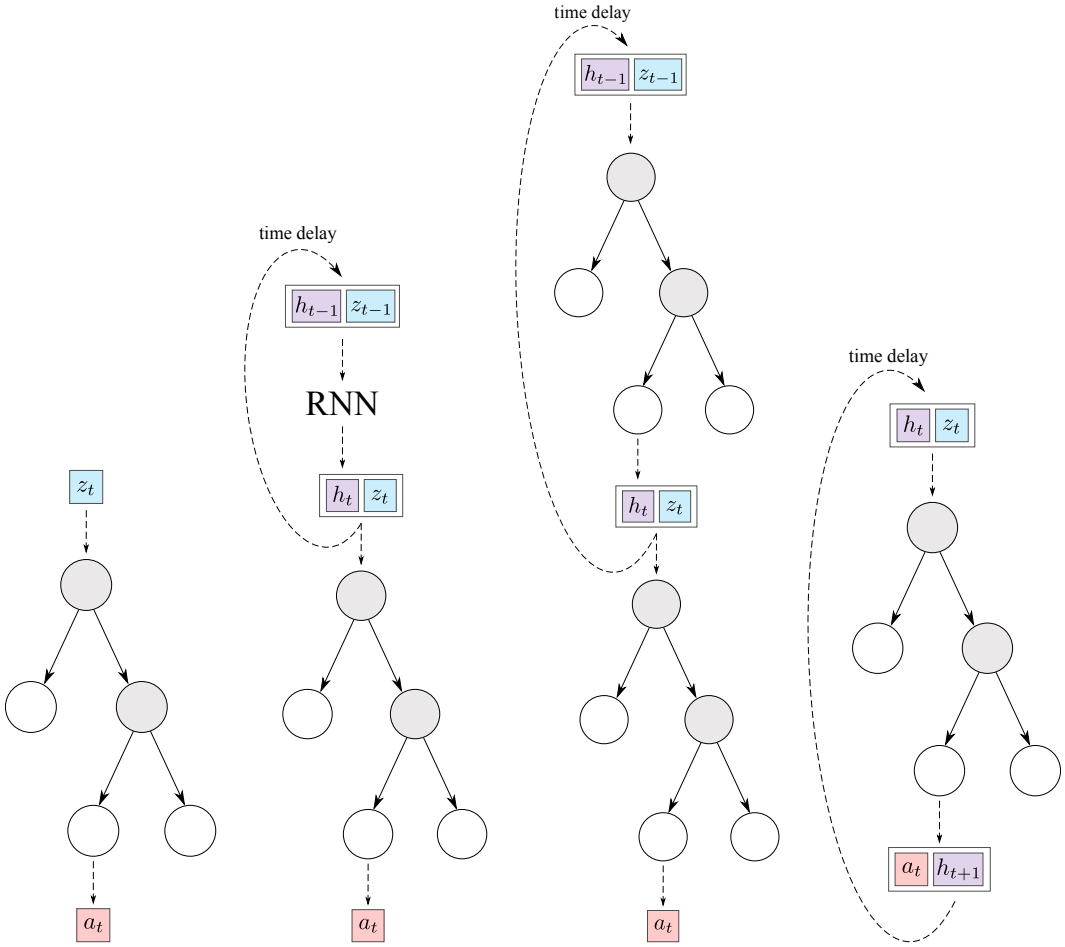
Validation performance in Algorithm 1 can be determined with traditional supervised learning metrics, such as target prediction accuracy, as well as with alternative success measures such as interpretability ratings or domain-expert approval.

A particular implementation challenge is to minimise the loss of information from previous optimisations during growth of the history tree, as new leaves are created (step 2a in Figure 8b). To

overcome this, split leaves are initialised to the previous value of the parent node, with added Gaussian noise of variance $\sigma^2 = 1/d$ where d is the leaf depth. This facilitates local optimisation as the tree grows, by ensuring new leaves still capture similar distributions to optimised parent nodes.

B.1 Interpretable representation learning for time-series

Our method relies on history representation learning for partially-observable environments: this section demonstrates the gains achieved by our choice of model. In particular, our Recurrent Decision Tree (RDT) models address our requirement for *fully-interpretable policies*. Different structures shown in Figures 6b and 6c were also evaluated, in which h_t is first extracted from $z_{1:t-1}$ through an RNN or RDT, and is concatenated to latest observation z_t as input to a distinct action-decision tree. Results in Table 4 highlight that our RDT architecture performs as well as the RNN history extractor in action-matching performance, and, as expected, is preferred by surveyed physicians for its interpretability. Performance remained acceptable as history dimensionality was decreased to 1 ($< 5\%$ loss in accuracy), which is valuable for visualisation purposes.



(a) DT policy. (b) RNN+DT policy. (c) RDT+DT policy. (d) RDT policy.

Figure 6: **Decision tree policies with different history extraction models**.

The surprising fact that our decision trees perform history representation nearly as well as neural networks can be attributed to the sparse and heterogeneous nature of our tabular data, often better captured through models based on decision trees (Frosst & Hinton, 2017). It is however not expected that RDTs would be applicable to complex, structured history embeddings as in speech or language modelling.

Leaf recurrence models An important step in model design consists of choosing leaf parameters for history prediction. In increasing order of complexity in terms of numbers of parameters per leaf,

Table 4: **Performance of different history extraction models for decision tree policies on ADNI.** Unless shown, standard errors were all ≤ 0.02. History-extraction and action-prediction trees have respective depths d_H and d_A.

History extraction model	d_H	d_A	# Parameters	Interpretability	AUPRC	Brier
None (DT)	–	3.7 ± 0.4	36 ± 4	9.3 ± 0.3	0.72	0.25
RNN+DT	–	4.8 ± 0.2	253 ± 6	5.0 ± 0.5	0.83	0.17
RDT+DT	3.3 ± 0.4	2.5 ± 0.3	167 ± 12	6.3 ± 0.3	0.80	0.21
RDT	3.3 ± 0.7		122 ± 18	8.3 ± 0.3	0.82	0.18

Table 5 highlight a number of possibilities, where $\{b \in \mathbb{R}^m,\ W \in \mathbb{R}^{m \times m},\ U \in \mathbb{R}^{m \times d},\ w \in \mathbb{R}^m\}$ are trainable parameters for each leaf (for history and observation dimensionality m and d respectively) and $\odot$ denotes element-wise multiplication. As for action prediction, the simplest solution is for leaves to be independent of inputs, outputting a fixed history vector – either a distribution over categorical values with equation 9, or a vector of continuous values through equation 10. Dependence on previous history h_{t-1} and on the newest observation z_{t-1} can be incorporated through element-wise or matrix multiplications (equations 11–15). In particular, equation 15 corresponds to an RNN unit.

Table 5: **Performance of different leaf models for history recurrence on ADNI**, for hidden unit dimension 8. Standard error on tree depth d is omitted for brevity.

Recurrence equation		d	# Parameters	Accuracy	Warm Start Acc.
$h_{t+1} = \text{softmax}(b)$	(9)	2.7	114 ± 11	0.72 ± 0.02	–
$h_{t+1} = \tanh(b)$	(10)	3.0	130 ± 23	0.76 ± 0.03	–
$h_{t+1} = \tanh(b + w \odot h_t)$	(11)	2.7	154 ± 26	0.68 ± 0.01	–
$h_{t+1} = \tanh(b + Uz_t)$	(12)	3.0	335 ± 46	0.76 ± 0.01	0.73 ± 0.01
$h_{t+1} = \tanh(b + w \odot h_t + Uz_t)$	(13)	2.3	306 ± 19	0.75 ± 0.02	0.73 ± 0.01
$h_{t+1} = \tanh(b + Wh_t)$	(14)	3.3	522 ± 110	0.73 ± 0.02	0.77 ± 0.02
$h_{t+1} = \tanh(b + Wh_t + Uz_t)$	(15)	3.7	847 ± 152	0.76 ± 0.02	0.77 ± 0.01

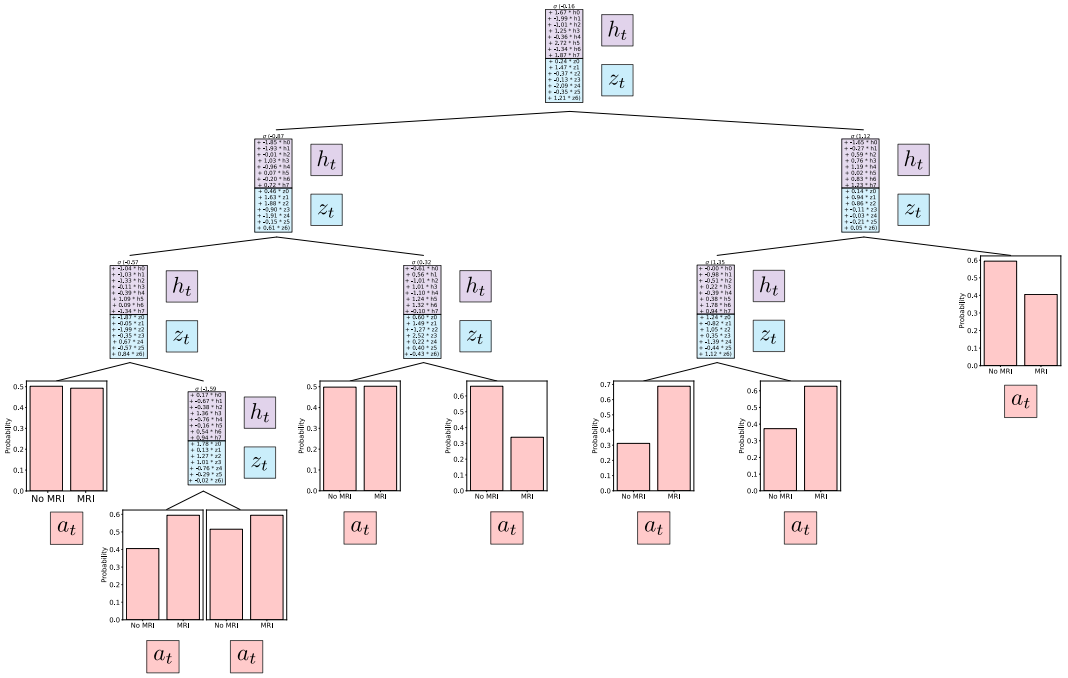
More complex leaf models tend perform best on ADNI, at the cost of tree readability and computation cost. Still, using fixed leaf history parameters through a hyperbolic tangent activation (equation 10) result in second-best performance with a minimal number of parameters, and intuitive readability – justifying the use of this parsimonious model for all other investigations. Initialising parameters with values previously optimised for the RNN+DT recurrence model, in a *warm start*, was found to achieve similar performance, but did not systematically improve results. A caveat should be added, as some decision-making environments require more complex recurrence models to capture subtleties in patient history. For instance, policy optimisation for our proof-of-concept SYNTH dataset was challenging with history-independent models, and recurrence equation 14 was used instead. Overall, this architecture choice depends on the complexity of the modelling task and on the nature of demonstrated trajectories.

B.2 Improvement on established decision tree models

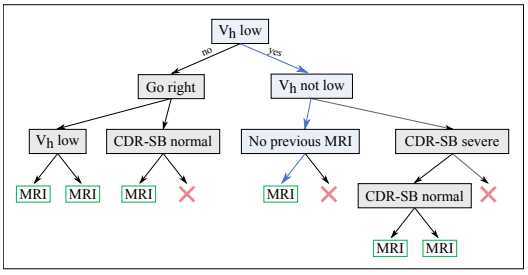
We conclude this section with a comparison of our method with the traditional Classification And Regression Tree (CART) and SDT algorithms in Table 6, to highlight our contributions. In particular, our work combines soft and grown tree architectures (Frosst & Hinton, 2017; Tanno et al., 2019) and pioneers their application to tabular data instead of images. We formalise and study the conversion of multidimensional gating functions to axis-aligned ones for interpretability. Our final contribution is an entirely novel approach to integrate time-series representation learning through a multi-output recurrent structure.

C Axis-aligned tree approximation

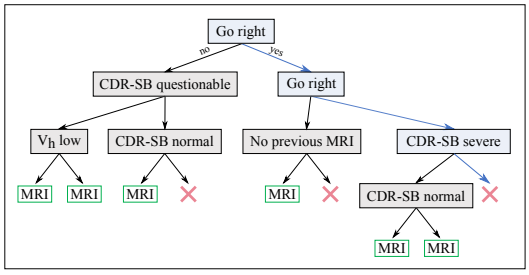
A final technical consideration required by our model architecture is policy post-processing for human interpretation. After optimisation, our decision trees consist of multidimensional gating functions over the history and observations variables which are difficult to understand – as illustrated in Figure 7a, and must be converted to axis-aligned, unidimensional thresholds.



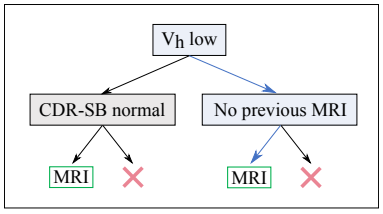
(a) Multidimensional policy.



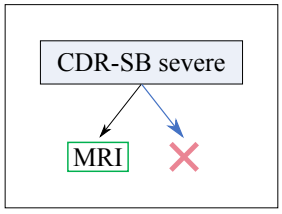
(b) Axis-aligned policy, $t = 1$.



(c) Axis-aligned policy, $t = 4$.



(d) Axis-aligned policy after pruning, $t = 1$.



(e) Axis-aligned policy after pruning, $t = 4$.

Figure 7: **Transformation of multidimensional tree policy into axis-aligned and pruned structure.** History and expected outcome outputs are not represented for clarity.

Table 6: **Comparison of our proposed architecture with traditional and soft decision trees (SDT)** as in Breiman et al. (1984) and Frosst & Hinton (2017).

	CART	SDT	**PoETree (Ours)**
MODELLING TASKS			
Discrete, categorical, continuous outputs	✓	✓	✓
Multiple outputs (e.g., $\{a_t, h_t, \tilde{z}_t\}$)	✗	✗	✓
Multidimensional decision boundaries	✗	✓	✓
Intepretable decision boundaries	✓	✗	✓
Time-dependence handling via recurrence	✗	✗	✓
OPTIMISATION			
Optimisation objective	Gini impurity	Prediction error	Prediction error
Probabilistic decision boundaries	✗	✓	✓
Gradient-descent optimisation	✗	✓	✓
Tree depth growth & path pruning	✗	✗	✓

Each inner node thus consists of learned parameters $\{b, \mathbf{w}\}$ and computes the probability of taking the right-most branch as follows:

$$p_{gate}(h_t, z_t) = \sigma\left(b + \sum_{i=1}^{m} w_i \cdot h_{t,i} + \sum_{i=1}^{d} w_{m+i} \cdot z_{t,i}\right) \tag{16}$$

where m and d are history and observation space dimensionalities.

The history vector is first marginalised out into the bias term, to condition the resulting observation-to-action policy on the history as desired. This results in a specific tree model for each datapoint, which only depends on observation z_t for easier human interpretation. Gating functions are re-expressed as follows:

$$p_{gate}(z_t) = \sigma\left(b' + \sum_{i=1}^{d} w_{m+i} \cdot z_{t,i}\right) \tag{17}$$

where $b' = b + \sum_{i=1}^{m} w_i \cdot h_{t,i}$ is now a history-dependent bias term.

As in Silva et al. (2020), these time-dependent multidimensional policies are then converted into axis-aligned unidimensional trees by taking the largest component of the remaining gating parameters at each node, w_{max}:

$$p_{gate}(z_t) > 0.5 \quad \approxeq \quad b' + w_{max} \cdot z_{t,max} > 0, \quad \text{where } w_{max} = \max\{w_{m+i}\}_{i=1}^{d} \tag{18}$$

$$\implies z_{t,max} > -b'/w_{max} \tag{19}$$

$$(\text{or} \quad z_{t,max} < -b'/w_{max} \quad \text{if } w_{max} < 0) \tag{20}$$

Figures 7b and 7c are examples of axis-aligned policies at different timesteps for typical patient B (with degrading MCI symptoms), much easier to follow than the multidimensional tree in Figure 7a.

Table 7: **Performance of multidimensional and different axis-aligned DT policy structures.** Standard errors all $\leq 1\%$.

Adaptive DT Policy	ADNI Accuracy (%)	MIMIC Accuracy (%)
Multidimensional	77.6	79.9
Axis-aligned; bias b'	74.7	75.9
Axis-aligned; bias $b' + \sum_{i \neq max} w_i \cdot z_{t,i}$	76.9	77.8
Axis-aligned; bias $b' + \sum_{i \neq max} w_i \cdot \tilde{z}_{t,i}$	75.4	76.0

Axis-aligned thresholds were also improved by exploiting our framework to predict expected observations at the subsequent timestep $\tilde{z}_{t+1}$. Thresholds were adjusted by *incorporating the predictions* for all components except $\tilde{z}_{t,max}$ in the bias term, transforming equation 19 into:

$$z_{t,max} > -\frac{b' + \sum_{i \neq max} w_i \cdot \tilde{z}_{t,i}}{w_{max}} \tag{21}$$

As shown in Table 7, incorporating true and expected observation values in the thresholds improved the axis-aligned model performance. Naturally, true test observations for the corresponding timestep

are unavailable, and can therefore not be used to adjust the threshold; only expected observations can be used for this. This investigation highlights another benefit of the expected patient outcome model, in addition to explanatory insights.

Further post-processing of axis-aligned policies in Figures 7b and 7c is required in the form of *node pruning*, setting a minimum path probability threshold on the validation set, to address the following concerns:

- Axis-aligned thresholds are often sensible and typically in the $[0, 1]$ range for categorical variables, converting straightforwardly to meaningful decisions. Negative or out-of-scope thresholds, however, exclude all data from a child path (e.g., left branch of the "Go right" node in Figure 7b), which can be pruned from the tree.
- Learning interdependencies between variables within the hierarchical structure is also challenging. For example, in Figure 7b, the first right branch (low V_h on previous scan) is incompatible with the following right branch (*not* low V_h). In practice, we obtained this structure due to 1) soft partitions, with examples shared across incompatible branches; and as 2) multi-dimensional gating affects thresholds in training. Again, the solution is to eliminate non-sensical children nodes in post-processing.

Accuracy loss was found to be virtually null ($< 2\%$) with a pruning probability threshold of 0.05 on the validation set $\mathcal{D}_{val}$. Formally, average path probability for node l was estimated as:

$$P^l(\mathcal{D}_{val}) = \frac{1}{|\mathcal{D}_{val}|} \sum_{\{h_t, z_t\} \in \mathcal{D}_{val}} P^l(h_t, z_t) \tag{22}$$

and nodes with $P^l(\mathcal{D}_{val}) < 0.05$ were eliminated from the structure.

Axis-aligned trees offered excellent performance, comparable to multidimensional ones. As the observation space dimensionality d increases, however, approximations become less accurate. To overcome this, trees can be trained closer to their axis-aligned version during optimisation, for instance with L1 regularisation to induce sparsity in the gating functions. On a 15-dimensional observation space with the SYNTH simulated policy, an L1-regularisation weight of 0.01 on gating parameters decreased multidimensional action-matching accuracy by 0.7% overall, but improved axis-aligned performance by 5%. This suggests a promising avenue of further work to ensure the scalability of our model.

D ALGORITHM IMPLEMENTATION DETAILS

Our PoETree algorithm was implemented as follows for our experimental investigations. Models were built using the open-source automatic differentiation framework JAX[3].

History representation learning was integrated in the decision, through the recurrent decision tree structure described in Section 3.2. Leaf models for history prediction were independent of previous history: $h_{t+1} = \tanh(b)$, where $b \in \mathbb{R}^m$ is a trainable parameter for each leaf and m is the history dimension. The history-to-action classifier π_h for causal regularisation was implemented as a simple fully-connected network with ReLU activation.

Following Algorithm 1, tree structures were initialised from depth 2, to avoid excessively simple local optima. For tree growth, limited to depth 5, the AUROC score was used as a measure of validation performance, to determine whether to split leaves into new inner nodes. All global and local optimisations were carried out until convergence – no improvements on the validation set for 50 update iterations. We used the Adam optimiser (Kingma & Ba, 2015) with a step size of 0.001 and a batch size of 32. Model hyperparameters in Table 8 were optimised through grid search on validation datasets – random subsets of 10% of the training data.

All experiments were run with 5 different model initialisations and data splits and the average results and standard errors were reported. Optimisation of tree structures was found to be temperamental, with some initialisations not improving in performance with training. A restarting procedure was therefore implemented if loss did not decrease by more than 5% over the first five epochs; and was

[3] `https://github.com/google/jax`

Table 8: **Hyperparameter grid for PoETree optimisation**, with values optimised for the ADNI dataset.

Hyperparameter	Search range	Optimal values
History dimension m	1 – 20	8
Splitting penalty λ	10^{-4} – 10^{1}	10^{-1}
Outcome prediction δ_1	10^{-3} – 10^{-1}	10^{-2}
Outcome prediction δ_2	10^{-4} – 10^{-1}	10^{-3}

also applied to benchmark algorithms if necessary. Experiments were performed on a Microsoft Azure virtual machine with 6 cores and powered by a Tesla K80 GPU.

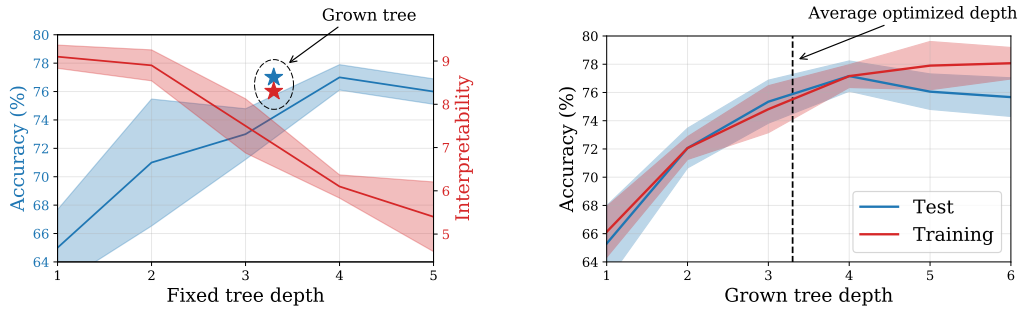
Complexity Analysis This section highlights the benefit of our tree growth procedure for both interpretability, complexity and performance optimisation.

Table 9: **Complexity analysis of different policy learning algorithms.**

Model	Optimisation time (s)	Runtime (s)
PoETree		
PoETree with inductive bias (Figure 10)		
Interp. PO-BC-IL (Hüyük et al., 2021)		

RUN AND COMMENT TIME ANALYSIS.

Figure 8a compares results obtained from trees of fixed depth and trees grown during training. Incrementally grown trees outperformed all fixed structures with a minimal number of parameters (with, on average, 40% fewer parameters than a fixed tree of the same depth) and reasonable training time (faster than fixed trees of depth 4 or 5, despite multiple local optimisations). Trees grown during the training process thus learn to capture general hierarchies in the data while eliminating unnecessary partitions, which otherwise make the tree more difficult to read as depth increases (Lage et al., 2018).



(a) Fixed trees

(b) Grown trees

Figure 8: **Performance on ADNI as a function of tree depth, for fixed and grown trees.** Our growth optimisation procedure adapts model complexity to the task at hand to (a) facilitate interpretation and (b) avoid overfitting.

A comparison of performance against tree depth after optimisation is also proposed in Figure 8b, demonstrating that our incrementally grown structure adapts its complexity to the the task at hand. Indeed, while performance on the training set continuously increases with tree depth, test accuracy is optimal at intermediate model complexities which avoid overfitting while capturing subtleties in the data. With an average tree depth of 3.3 ± 0.7, our optimisation procedure therefore successfully implements this Occam's razor.

ADD SAMPLE COMPLEXITY analysis

E BENCHMARKS IMPLEMENTATION DETAILS

IOHMM parameters for all learned environment models are initialised uniformly at random.

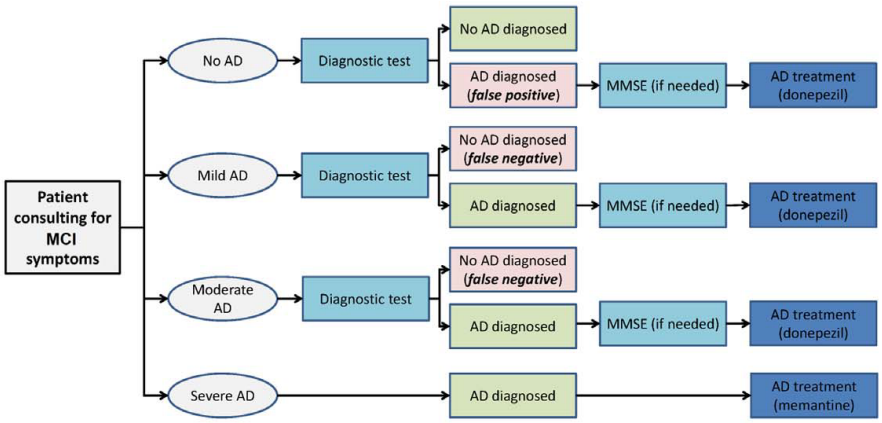


Figure 9: **Published clinical diagnostic policy for patient with MCI symptoms.** Figure reproduced from Biasutti et al. (2012).

PO-IRL After training and freezing IOHMM parameters (Bengio & Frasconi, 1995), the reward function is initialised as $R(s, a) \sim \mathcal{N}(0, 0.001^2)$ and estimated by averaging over 50 Markov Chain Monte Carlo (MCMC) samples – taking every tenth value after a burn-in of 500 samples. In **Offline PO-IRL**, IOHMM parameters and reward signal are jointly learned by MCMC sampling. Q-values are then recovered with an off-the-shelf PO-MDP solver[4].

PO-MB-IL IOHMM parameters are first trained (Bengio & Frasconi, 1995) and freezed. Policies are parametrised in terms of decision boundaries as in Hüyük et al. (2021), and initialised randomly. Training is carried out by maximising the likelihood of actions in the demonstrations dataset by Expectation-Maximisation. Maximisation is carried out with the Adam optimizer (Kingma & Ba, 2015) with learning rate 0.001 until convergence.

F Dataset details

F.1 Proof-of-concept policy

SYNTH dataset: Efficacy on diseased patients is 40%, with no spontaneous state changes and an initial 80% diseased population. Diagnostic tests have 99% precision and 5% false alarms.

F.2 The Alzheimer's Disease Neuroimaging Initiative (ADNI)

Details on ADNI: Average hippocampal volumes were taken as within half a standard deviation from the population mean.

Published guidelines for Alzheimer's disease investigation are reproduced in Figure 9. With the help of one of the surveyed physicians, this was distilled as a static tree policy in Figure 10 to initialise PoETree and assess the effect of inductive bias on optimisation performance. Gaussian noise of variance $\sigma^2 = 0.1$ was added to initialisation parameters. Table 9 evidences the gains afforded by this incorporation of prior knowledge.

F.3 Medical Information Mart for Intensive Care (MIMIC-III)

Policies recovered for the MIMIC antibiotic prediction task by our algorithm and the decision-boundary framework (Hüyük et al., 2021) are given in Figure 11. The latter again falls short of highlighting how patients are mapped onto the belief state, giving our decision tree greater expressive power. Our results are consistent with policy learning work on the same task and dataset (Bica et al., 2021), which identifies similar discriminative features: fever and abnormally high or low white blood cell counts are known clinical criteria for infection and antibiotic treatment, if a bacterial source is suspected (Masterton et al., 2008). Continuous variable thresholds evolve over time through their dependence on history: the temperature criterion, for instance, adjusts to whether the patient has a prior for infection.

[4] `https://www.pomdp.org/code/index.html`.

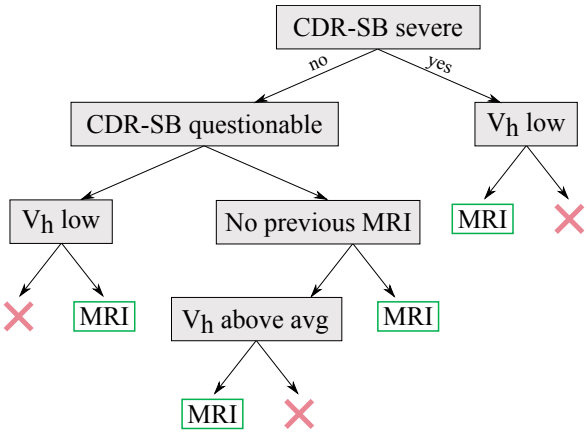
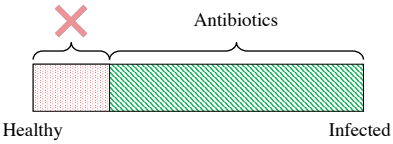
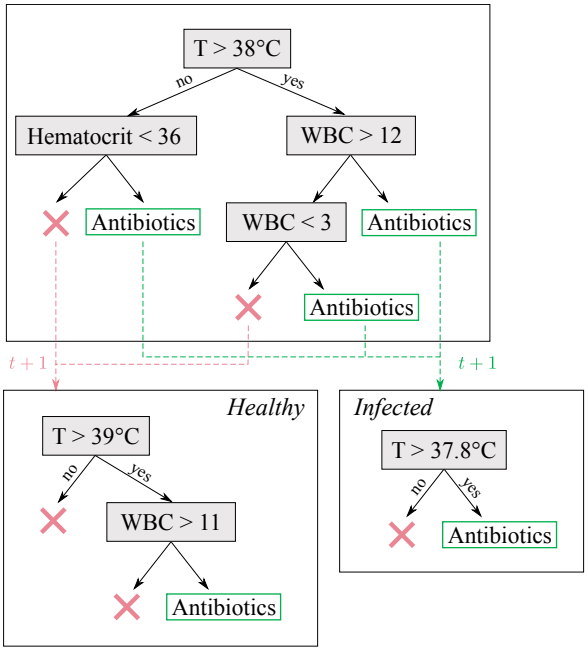


Figure 10: **Policy initialisation for inductive bias analysis**, based on Figure 9.



(a) Decision-boundary policy (Hüyük et al., 2021).



(b) Adaptive decision tree policy (only 2 timesteps visualised).

Figure 11: **Recovered policies for MIMIC.** Units for white blood cell count (WBC) are 1000/mm^3.

G DISCUSSION WITH CLINICIANS

To evaluate the interpretability of our decision tree policies, we consulted three practising physicians for their feedback on our work in comparison to the state-of-the-art. Each physician was given a short presentation introducing (1) our goal of representing the decision-making behaviour in an intelligible way, (2) the ADNI dataset and the studied clinical environment, (3) how policies can be represented in terms of decision boundaries over belief simplices (Hüyük et al., 2021), black-box mappings from beliefs to actions (Sun et al., 2017), and direct mappings from observations to actions (as in our work). All participants were unaware of what method we proposed, to minimise any implicit bias in results.

All three clinicians expressed the following preference in terms of interpretability of the decision-making process:

$$\text{RNN} < \text{Decision Boundaries} < \text{Decision Trees}$$

where the RNN policies, learned by PO-BC-IL (Sun et al., 2017), correspond to black-box mappings from observation to history, and from history to actions. This was illustrated in our survey through Table 10. Decision-boundary policies correspond to the interpretable PO-BC-IL benchmark (Hüyük et al., 2021).Finally, decision tree policies learned through our PoETree framework were illustrated by Figure 2a.

Although more participants would be required for greater reliability, this survey was useful to confirm end-user's preference for our representation of behaviour through trees. Additional comments on the belief space and decision-boundary policy (Hüyük et al., 2021) are summarised below:

- Classifying patients within discrete disease states aligns with the cognitive process followed in clinical practice (O'Bryant et al., 2008), but is challenging to achieve in practice, as patients often fall on a **continuum of symptoms and illness severity**.
- **Visual representations** of belief states can be misleading. In particular, the triangular representation in Hüyük et al. (2021) does not evidence the linear progression from normal to MCI to dementia, considered the traditional patient evolution (NICE, 2018). In addition, higher-dimensional, hierarchical or overlapping belief states cannot be readily visualised.
- The mapping from **patient observations $\mathcal{Z}$ to belief space $\mathcal{S}$**, and the **evolution of patient trajectories** in this space as new information is acquired, is unclear.

Feedback on our decision tree policies was also collected, highlighting important advantages and limitations:

- The flow chart captured by the decision tree was described by clinicians as most similar to their own approach to care, hierarchically eliminating possible diagnoses, treatments or investigations – in agreement with cognitive research (Zylberberg et al., 2017) and established guidelines illustrated in Figure 9.
- The concept of **action value** (for diagnostic tests, value of information; for treatments, patient improvement) is captured in an easily understandable way in terms of expected patient evolution. In contrast, they noted, reward functions and belief updates proposed by related work (Makino & Takeuchi, 2012; Hüyük et al., 2021) do not represent the medical thought-process as faithfully.
- Although the recurrent decision tree mapping from leaves to a subsequent tree policy was appreciated, there remain limitations in terms of the **interpretability of the history embedding**, particularly with complex leaf recurrence models. In further work, we could visualise history embeddings over a belief simplex following Hüyük et al. (2021), conditioning vertices to encode distinct beliefs about the patient condition (e.g., diagnoses). Different regions of belief space would thus be associated with different tree policies, rather than a single most-likely action.

Overall, feedback from our domain experts and intended end-users was useful to validate the priorities of our work. In particular, their strong preference for transparent representations of decision-making motivated an in-depth study of how to best achieve this within our system, guided by interpretability and sensibility ratings.

Table 10: **Patient trajectories under a policy learned by PO-BC-IL**. History embeddings are obtained through a recurrent neural network.

Timestep	Observation z_t		History h_t	Action a_t
	CDR-SB	Previous MRI		
$t = 1$	Normal	Average V_h	$[+0.5, -0.5, +0.5, +1.0, +0.2, -0.1, +0.9, -0.9]$	✗
$t = 2$	Normal	✗	$[-0.4, +1.0, -0.7, +0.9, +0.8, +1.0, -0.9, +0.6]$	✗
$t = 3$	Normal	✗	$[-0.2, +0.1, -0.1, +0.6, +0.9, +0.8, +0.6, -0.4]$	✗
$t = 4$	Normal	✗	$[-0.2, +0.8, -0.6, +0.3, +0.8, +1.0, +0.0, +0.4]$	✗

(a) Patient A (Healthy)

Timestep	Observation z_t		History h_t	Action a_t
	CDR-SB	Previous MRI		
$t = 1$	Questionable	Low V_h	$[+0.8, +0.3, +0.3, +0.2, +0.1, +0.7, -0.5, -0.7]$	MRI
$t = 2$	Questionable	Low V_h	$[+0.9, -0.1, -0.2, +0.6, -0.2, +0.1, -0.5, -0.8]$	MRI
$t = 3$	Questionable	Low V_h	$[+0.9, +0.6, -0.4, +0.9, -0.1, +0.6, -0.7, -0.7]$	MRI
$t = 4$	Severe	Low V_h	$[+0.8, +0.2, -0.9, +0.9, -0.5, -0.6, -0.8, -0.8]$	✗

(b) Patient B (Degrading)

Timestep	Observation z_t		History h_t	Action a_t
	CDR-SB	Previous MRI		
$t = 1$	Severe	Low V_h	$[-0.5, +0.9, +0.2, +0.5, +0.3, +0.5, -0.9, -0.9]$	✗
$t = 2$	Severe	✗	$[-0.7, -0.5, +0.8, -0.4, +0.3, -0.1, -0.5, +0.7]$	✗
$t = 3$	Severe	✗	$[-0.8, +0.7, +0.9, -0.8, +0.6, -0.1, -0.9, +0.9]$	✗
$t = 4$	Severe	✗	$[-0.8, +0.4, +1.0, -0.9, +0.7, -0.4, -0.9, +0.9]$	✗

(c) Patient C (Dementia)