Appendices

Appendix A gives a longer discussion of merits and caveats; Appendix B gives further experiment details; Appendix C gives derivations of propositions; Appendix D shows illustrative trajectories; Appendix E gives a summary of notation.

A. Discussion

In this paper, we motivated the importance of descriptive models of behavior as the bridge between normative and prescriptive decision analysis [9–11] (Figure 4). On account of this, we formalized a unifying perspective on inverse decision modeling for behavior representation learning. Precisely, the inverse decision model of any observed behavior \( \phi_{\text{demo}} \) is given by its projection \( \phi_{\text{imit}}^* = F_{\Theta_{\text{norm}}} \circ G_{\Theta_{\text{norm}}} (\phi_{\text{demo}}) \) onto the space \( \Phi_{\Theta_{\text{norm}}} \) of behaviors parameterizable by the structure designed for \( \Theta \) and normative standards \( \Theta_{\text{norm}} \) specified. This formulation is general. For instance, it is agnostic as to the nature of agent and environment state spaces (which—among other properties—are encoded in \( \psi \)); it is also agnostic as to whether the underlying forward problem is model-free or model-based (which—among other properties—is encoded in \( \theta \)). Per the priorities of the investigator (cf. imitation, apprenticeship, understanding, and other objectives), different choices can and should be made to balance the expressivity, interpretability, and tractability of learned models.

Partial Observability At first glance, our choice to accommodate partial observability may have appeared inconsequential. However, its significance becomes immediately apparent once we view an agent’s behavior as induced by both a decision policy \( \pi \) as well as a recognition policy \( \rho \), and importantly—that not only may an agent’s mapping from internal states into actions be suboptimal (viz. the former), but that their mapping from observations into beliefs may also be subjective (viz. the latter). Therefore in addition to the oft-studied, purely utility-centric nature of (perfectly rational) behavior, this generalized formalism immediately invites consideration of (boundedly rational) behaviors—that is, agents acting under knowledge uncertainty, biased by optimism/robustness, with policies distorted by the complexities of information processing required for decision-making.

Bounded Rationality While the IDM formalism subsumes most standard approaches to imitation learning, apprenticeship learning, and reward learning (cf. Table 1 and Table 3), we emphasize that—with very few exceptions [78–80]—the vast majority of original studies in these areas are limited to cases where \( \theta_{\text{desc}} = v \) alone, or assume fully-observable environments (whence \( S = X = Z \), and \( \rho \) simply being the identity function). Therefore our concrete example of inverse bounded rational control was presented as a prototypical instantiation of IDM that much more fully exercises the flexibility afforded by this generalized perspective. Importantly, while our notion of bounded rationality has (implicitly) been present to varying degrees in (forward) control and reinforcement learning (cf. Table 2 and Table 4), “boundedness” has largely been limited to mean “noisy actions”. To be precise, we may differentiate between three “levels” of boundedness:

- **Imperfect Response**: This is the shallowest form of boundedness, and includes Boltzmann-exploratory [142–144] and (locally) entropy-regularized [170] behaviors: It considers first that agents are perfect in their ability to compute the optimal values/policies; however, their actions are ultimately executed with an artificial layer of stochasticity.
- **Capacity Constraints**: Given an agent’s model (e.g. \( \tau \), Q-network, etc.), the information processing needed in computing actions on the go is costly. We may view soft-optimal [101–104] and KL-regularized [107–111] planning and learning as examples. However, these do not model subjectivity of beliefs, adaptivity, or optimism/robustness.
- **Model Imperfection**: The agent’s mental model itself is systematically flawed, due to uncertainty in knowledge, and to biases from optimism or pessimism. We may view certain robust MDPs (with penalties for deviating from priors) [148–151] as examples. However, these still do not account for partial observability (and biased recognition).

Now in the inverse direction, imitation/apprenticeship learning has typically viewed reward learning as but an intermediary, so classical methods have worked with perfectly rational planners [13, 40–45, 70–73]. Approaches that leverage probabilistic methods have usually simply used Boltzmann-exploratory policies on top of optimal action-value functions (viz. imperfect response) [49–52, 59–63, 74, 75], or worked within maximum entropy planning/learning frameworks
(viz. capacity constraints) [81–92]. Crucially, however, the corresponding parameters (i.e. inverse temperatures) have largely been treated as pre-specified parameters for learning $v$ alone—not learnable parameters of interest by themselves. In contrast, what IDM allows (and what IBRC illustrates) is the “fullest” extent of boundedness—that is, where stochastic actions and subjective beliefs are endogenously the result of knowledge uncertainty and information processing constraints. Importantly, while recent work in imitation/apprenticeship have studied aspects of subjective dynamics that can be jointly learnable [67–69, 93, 94], they are limited to environments that are fully-observable and/or agents that have point-valued knowledge of environments—substantial simplifications that ignore how humans can and do make imperfect inferences from recognizing environment signals.

A.1. Important Distinctions

Our goal of understanding in IDM departs from the standard objectives of imitation and apprenticeship learning. As a result, some caveats and distinctions warrant special attention as pertains assumptions, subjectivity, and model accuracy.

Decision-maker vs. Investigator As noted in Section 3.3, the design of $\Theta$ (and specification of $\theta_{\text{norm}}$) are not assumptions: We are not making “factual” claims concerning the underlying psychological processes that govern human behavior; these are hugely complex, and are the preserve of neuroscience and biology [171]. Instead, such specifications are active design choices: We seek to make the “effective” claim that an agent is behaving as if their generative mechanism was parameterized by the (interpretable) structure we designed for $\Theta$. Therefore when we speak of “assumptions”, it is important to distinguish between assumptions about the agent (of which we make none), versus assumptions about the investigator performing IDM (of which, by construction, we assume they have the ability to specify values for $\theta_{\text{norm}}$).

In IBRC, for example, in learning $\beta$ we are asking the question: “How much (optimistic/pessimistic) deviation from neutral knowledge does the agent appear to tolerate?” For this question to be meaningfully answered, we—as the investigator—must be able to produce a meaningful value for $\hat{\sigma}$ to specify as part of $\theta_{\text{norm}}$. In most cases, we are interested in deviations from some notion of “current medical knowledge”, or what knowledge an “ideal” clinician may be expected to possess; thus we may—for instance—supply a value for $\hat{\sigma}$ via models learned a priori from data. Of course, coming up such values for $\theta_{\text{norm}}$ is not trivial (not to mention entirely dependent on the problem and the investigator’s objectives regarding interpretability); however, we emphasize that this does not involve assumptions regarding the agent.

Subjective vs. Objective Dynamics In imitation and apprenticeship learning, parameterizations of utilities and dynamics models are simply intermediaries for the downstream task (of replicating expert actions or matching expert returns). As a result, no distinction needs be made between the “external” environment (with objective dynamics $\tau_{\text{env}}, \omega_{\text{env}}$) and the “internal” environment model that an agent works with (with subjective dynamics $\tau, \omega$). Indeed, if the learned model were to be evaluated based on live deployment in the real environment (as is the case in IL/IRL), it only makes sense that we stipulate $\tau, \omega = \tau_{\text{env}}, \omega_{\text{env}}$ for the best results.

However, in IDM (and IBRC) we are precisely accounting for how an agent may appear to deviate from such perfect, point-valued knowledge of the environment. Disentangling subjective and objective dynamics is now critical: Both the forward recursion (Lemma 1) for occupancy measures and the backward recursion (Theorem 4) for value functions are computations internal to the agent’s mind—and need not correspond to any notion of true environment dynamics. The external dynamics only comes into play when considering the distribution of trajectories $h \sim \phi_{\tau, \omega}$ induced by an agent’s policies, which—by definition—manifests through (actual or potential) interaction with the real environment.

Demonstrated vs. Projected Behavior As advanced throughout, a primary benefit of the generalized perspective we develop is that we may ask normative-descriptive questions taking the form: “Given that this (boundedly rational) agent should optimize this $v$, how suboptimally do they appear to behave?” Precisely, as pertains IBRC we noted that—
as the investigator—we are free to specify (what we deem) “meaningful” values for \( v \) within \( \theta_{\text{norm}} \), while recovering one or more behavioral parameters \( \alpha, \beta, \gamma \) from \( \theta_{\text{desc}} \). Clearly, however, we are not at liberty to specify completely random values for \( v \) (or, more generally, that we are not at liberty to design \( \Theta \) and \( \theta_{\text{norm}} \) in an entirely arbitrary fashion). For one, the resulting inverse decision model may simply be a poor reflection the original behavior (i.e. the projection \( \phi^*_{\text{imit}} \) onto \( \Phi_{\theta_{\text{norm}}} \) may simply lose too much information from \( \phi_{\text{demo}} \).6

Without doubt, the usefulness of the inverse decision model (i.e. in providing valid interpretations of observed behavior) depends entirely on the design and specification of \( \Theta \) and \( \theta_{\text{norm}} \), which requires care in practice. Most importantly, it should be verified that—under our designed parameterization—the projected behavior \( \phi^*_{\text{imit}} \) is still a faithful model of the demonstrated behavior \( \phi_{\text{demo}} \). In particular, compared with fitting a black-box model for imitating behavior—or any standard method for imitation/apprenticeship learning, for that matter—it should be verified that our (interpretable parameterized) model does not suffer inordinately in terms of accuracy measures (i.e. in predicting \( u \) from \( h \); otherwise the model (and its interpretation) would not be meaningful.

In Appendix B, we perform precisely such a sanity check for IBRC, using a variety of standard benchmarks (Table 5).

### A.2. Further Related Work

While relevant works have been noted throughout the manuscript, here we provide additional context for IDM and IBRC, and how notable techniques/frameworks relate to our work.

**Inverse Decision Modeling** Pertinent methods subsumed by our forward and inverse formalisms have been noted in Tables 2–3. In particular, techniques that can be formalized as instantiations of IDM are enumerated in Table 1. Broadly, for imitation learning these include behavioral cloning-like methods [14–21], as well as distribution-matching methods that directly match occupancy measures [23–39]; we defer to [12,100] for more thorough surveys. For apprenticeship learning by inverse reinforcement learning, these include classic maximum-margin methods based on feature expectations [13,40–45], maximum likelihood soft policy matching using Boltzmann-rational policies [51,52], maximum entropy policies [50,89–92], and Bayesian maximum a posteriori inference [59–63], as well as methods that leverage preference models and annotations for learning [95–99].

In this context, the novelty of the IDM formalism is two-fold. First, in defining a unifying framework that generalizes all prior techniques, IDM simultaneously opens up a new class of problems in behavior representation learning with consciously designed parameterizations. Specifically, in defining inverse decision models as projections in \( \Phi \)-space induced by \( F, G, \) and \( \Theta \), the structure and decomposition chosen for \( \theta_{\text{norm}} \times \theta_{\text{desc}} \) allows asking normative-descriptive questions that seek to understand observed decision-making behavior. Second, in elevating recognition policies to first-class citizenship in partially-observable environments, IDM greatly generalizes the notion of “boundedness” in decision-making—that is, from the existing focus on noisy optimality in \( \pi \), to the ideas of subjective dynamics \( \sigma \) and biased belief-updates \( \rho \) (viz. discussion in the beginning of this section).

**Orthogonal Frameworks** Multiple studies have proposed frameworks that provide generalized treatments of different aspects of inverse reinforcement learning [28,30,35,58,60,172,173]. However, these are orthogonal to our purposes in the sense that they are primarily concerned with establishing connections between different aspects/subsets of the imitation/apprenticeship learning literature. These include loss-function perspectives [58] and Bayesian MAP perspectives [60] on inverse reinforcement learning, \( f \)-divergence minimization perspectives [28,30] on distribution matching, connections between adversarial and non-adversarial methods for distribution matching [35], as well as different problem settings for learning reward functions [173]. But relative to the IDM formalism, all such frameworks operate within the special case of \( \theta_{\text{desc}} = v \) (and full observability).

**Case Study: GAIL** Beyond aforementioned distinctions, another implication is that IDM defines a single language for understanding key results in such prior works. For example, we revisit the well-known result in [25] that gives rise to generative adversarial imitation learning (“GAIL”): It is instructive to recast it in more general—but simpler—terms.

First, consider a maximum entropy learner in the MDP setting (cf. Table 2), paired with a maximum margin identification strategy with a parameter regularizer \( \zeta \) (cf. Table 3):

\[
F_{\theta_{\text{norm}}}^\text{ME}(\theta_{\text{desc}}) = \phi_\pi \quad \text{where } \pi^* = \arg\max_{\pi} \mathbb{E}_{z \sim \rho_0} V_{\phi_\pi} \quad \text{(26)}
\]

Second, consider a black-box decision-rule policy (cf. Table 2), where neural-network weights \( \chi \) directly parameterize a policy network \( f_{\text{decision}} \) (and \( \theta_{\text{desc}} = \chi \)); this is paired with a distribution matching identification strategy (cf. Table 3):

\[
F_{\theta_{\text{norm}}}^\text{DR}(\theta_{\text{desc}}) = \arg\max_{\pi} \delta(\pi - f_{\text{decision}}(\chi)) \quad \text{(28)}
\]

where distance measures are given by the convex conjugate \( \zeta^* \), and \( H_{\text{limit}} \) gives the causal entropy of the imitating policy. Now, the primary motivation behind generative adversarial imitation learning is the observation that \( \zeta \)-regularized maximum-margin soft IRL implicitly seeks a policy whose occupancy is close to the demonstrator’s as measured by \( \zeta^* \). In IDM, this corresponds to a remarkably simple statement:
Proposition 6 (Ho and Ermon, Recast) Define the behavior projections induced by the composition of each pairing:

\[
\text{proj}^{\text{ME,MM}}_{\theta_{\text{term}}} = F^{\text{ME}}_{\theta_{\text{term}}} \circ G^{\text{MM}}_{\theta_{\text{term}}}
\]

(30)

\[
\text{proj}^{\text{DR,DM}}_{\theta_{\text{term}}} = F^{\text{DR}}_{\theta_{\text{term}}} \circ G^{\text{DM}}_{\theta_{\text{term}}}
\]

(31)

Then these projections are identical: \(\text{proj}^{\text{ME,MM}}_{\theta_{\text{term}}} = \text{proj}^{\text{DR,DM}}_{\theta_{\text{term}}}\) (and inverse decision models thereby obtained are identical).

In their original context, the significance of this lies in the fact that the first pairing explicitly requires parameterizations via reward functions (which—in classic apprenticeship methods—is restricted to be linear/convex), whereas the second pairing allows arbitrary parameterization by neural networks (which—while black-box—are more flexible). In our language, this simply means that the first projection requires \(\theta_{\text{desc}} = v\), while the second projection allows \(\theta_{\text{desc}} = \chi\).

**Inverse Bounded Rational Control** Pertaining to IBRC, methods that are comparable and/or subsumed have been noted in Tables 1 and 4. In addition, the context of IBRC within existing notions of bounded rationality have been discussed in detail in the beginning of this section. Now, more broadly, we note that the study of imperfect behaviors [4] spans multiple disciplines: in cognitive science [5], biological systems [6], behavioral economics [7], and information theory [8]. Specifically, IBRC generalizes this latter class of information-theoretic approaches to bounded rationality.

First, the notion of *flexibility* in terms of the informational effort in determining successive actions (cf. decision complexity) is present in maximum entropy [101–104] and KL-regularized [107–111] agents. Second, the notion of *tolerance* in terms of the statistical surprise in adapting to successive beliefs (cf. recognition complexity) is present in behavioral economics [7,174] and decision theory [75,146,175]. Third, the notions of *optimism* and *pessimism* in terms of the average regret in deviating from prior knowledge (cf. specification complexity) are present in robust planners [148–151].

On account of this, the novelty of the IBRC example is three-fold. First, it is the first to present generalized recursions incorporating all three notions of complexity—that is, in the mappings into internal states, models, and actions. Second, IBRC does so in the partially-observable setting, which—as noted in above discussions—crucially generalizes the idea of subjective dynamics into subjective beliefs, thereby accounting for boundedness in the recognition process itself. Third (perhaps most importantly), IBRC is the first to consider the inverse problem—that is, of turning the entire formalism on its head to learn the parameterizations of such boundedness, instead of simply assuming known parameters as required by the forward problem. Finally, it is important to note that IBRC is simply one example: There are of course many possibilities for formulating boundedness, including such aspects as myopia and temporal inconsistency [176,176]; we leave such applications for future work.

**Interpretable Behavior Representations** Lastly, a variety of works have approached the task of representing behaviors in an interpretable manner. In inverse reinforcement learning, multiple works have focused on the reward function itself, specifying interpretable structures that explicitly express a decision-maker’s preferences [62], behavior under time pressure [75], consideration of counterfactual outcomes [73], as well as intended goals [177]. Separately, another strand of research has focused on imposing interpretable structures onto policy functions themselves, such as representing policies in terms of decision trees [178] and intended outcomes [179] in the forward problem, or—in the inverse case—learning imitating policies based on decision trees [180] or decision boundaries [22]. In the context of IDM, both of these approaches can naturally be viewed as instantiations of our more general approach of learning representations of behavior through interpretable parameterized planners and inverse planners (as noted throughout Tables 1–3). Finally, for completeness also note that an orthogonal branch of research is dedicated to generating autonomous explanations of artificial behavior, as suggested updates to human models [181,182], and also as responses to human queries in a shared [183] or user-specified vocabulary [184].

**A.3. Future Work**

A clear source of potential research lies in exploring differently structured parameterizations \(\Theta\) to allow interpretable representation learning of behaviors. After all, beyond the black-box and reward-centric approaches in Table 1 and the handful of works that have sought to account for subjective dynamics [22,67,80,93], our example of IBRC is only one such prototype that exercises the IDM formalism more fully.

In developing more complex and/or expressive forward models, an important question to bear in mind is to what extent the inverse problem is identifiable. In most existing cases we have seen, the usual strategies—such as constraining scaling, shifting, reward shaping, as well as the use of Bayesian inference—is sufficient to recover meaningful values. However, we have also seen that in the extreme case of an arbitrary differentiable planner, any inverse problem immediately falls prey to the “no free lunch” result [105,106,136,137]. Thus balancing aspects of complexity, interpretability, and identifiability of decision models would be an interesting direction of work. Finally, in this work we primarily focused on the idea of limited intentionality—that is, in the goal-seeking nature of an agent and how they may be constrained in this respect. But the flip side is also interesting: One can explore the idea of limited attentionality—that is, in how an agent may be constrained in their ability to focus on sequences of past events. This idea is explored in [185,186] by analogy with information bottlenecks in sensors and memory capacities; however, there is much room for developing more human-interpretable parameterizations of how an agent may pay selective attention to observations over time.
**B. Experiment Details**

**Computation** In IBRC, we define the space of agent states (i.e. subjective beliefs) as $\mathcal{Z} = \mathbb{R}^k$, where $k$ is the number of world states ($k=3$ for ADNI, and $k=2$ for DIAG). To implement the backward recursion (Theorem 4), each dimension of $\mathcal{Z}$ is discretized with a resolution of 100, and the values $V(z)$ in the resulting lattice are updated iteratively exactly according to the backup operator $\mathbb{B}^*$—until convergence (which is guaranteed by the fact that $\mathbb{B}^*$ is contractive, therefore the fixed point is unique; see Appendix C). For evaluation at any point $z$, we (linearly) interpolate between the closest neighboring grid points. In terms of implementing the inverse problem in a Bayesian manner (i.e. to recover posterior distributions over $\Theta_{\text{desc}}$), we perform MCMC in log-parameter space (i.e. $\log \alpha, \log \beta, \log \eta$). Specifically, the proposal distribution is zero-mean Gaussian with standard deviation 0.1, with every 10th step collected as a sample. In each instance, the initial 1,000 burn-in samples are discarded, and a total of 10,000 steps are taken after burn-in.

**Recognition** In the manuscript, we make multiple references to the Bayes update, in particular within the context of our (possibly-biased) belief-update (Equation 9). For completeness, we state this explicitly: Given point-valued knowledge of $\tau, \omega$, update $p_{\tau,\omega}(z', u, x')$ is the Dirac delta centered at

$$p(s'|z, u, x', \tau, \omega) = \mathbb{P}_{\sim \tau}(z) \left[ \frac{\tau(s'|s, u)\omega(x'|u, s')}{\mathbb{E}_{\omega' \sim \tau(\cdot)}[\omega'(x'|u, s')]} \right]$$

(32)

and the overall recognition policy is the expectation over such values of $\tau, \omega$ (Equation 9). As noted in Section 4.1, in general $\tilde{\sigma}$ represents any prior distribution the agent is specified to have, and in particular can be some Bayesian posterior $p(\tau, \omega|\mathcal{E})$ given any form of experience $\mathcal{E}$. This can be modeled in any manner, and is not the focus of our work; what matters here is simply that the agent may deviate optimistically/pessimistically from such a prior. As noted in Section 5, for our purposes we simulate $\tilde{\sigma}$ by discretizing the space of models such that probabilities vary in ±10% increments from the (highest-likelihood) truth. In ADNI, this means $\tilde{\sigma}$ is centered at the IOHMM learned from the data.

**Model Accuracy** In Appendix A.1 we discussed the caveat: In order for an inverse decision model to provide valid interpretations of observed behavior, it should be verified that—under the designed parameterization—the projected behavior $\phi_{\text{mut}}^p$ is still an accurate model of the demonstrated behavior $\phi_{\text{demo}}$. Here we perform such a sanity check for our IBRC example using the ADNI environment. We consider the following standard benchmark algorithms. First, in terms of black-box models for imitation learning, we consider behavioral cloning [15] with a recurrent neural network for observation-action histories (RNN-Based BC-IL); an adaptation of model-based imitation learning [187] to partially-observable settings, using the learned IOHMM as model (IOHMM-Based BC-IL); and a recently-proposed model-based imitation learning that allows for subjective dynamics [22] by jointly learning the agent’s possibly-biased internal model and their probabilistic decision boundaries (Joint IOHMM-Based BC-IL). Second, in terms of classic reward-centric methods for apprenticeship learning, we consider Bayesian inverse reinforcement learning in partially-observable environments [75] equipped with the learned IOHMM as model (Bayesian PO-IRL); and—analogous to the black-box case—the equivalent of this method that trains the dynamics model jointly along with the agent’s apprenticeship policy [74] (Joint Bayesian PO-IRL). Algorithms requiring learned models are given IOHMMs estimated using conventional methods [188]—which is the same method by which the true model is estimated in IBRC (that is, as part of the space of candidate models in the support of $\tilde{\sigma}$).

**Results** Table 5 shows results of this comparison on predicting actions, computed using held-out samples based on 5-fold cross-validation. Crucially, while IBRC has the advantage in terms of interpretability of parameterization, its performance—purely in terms of predicting actions—does not degrade: IBRC is slightly better in terms of calibration, and similar in precision-recall (differences are statistically insignificant). Table 5. Comparison of Model Accuracies. IBRC performs similarly to all benchmark algorithms in matching demonstrated actions. Results are computed using held-out samples based on 5-fold cross-validation. IBRC is slightly better-calibrated, and similar in precision-recall scores (differences are statistically insignificant).

<table>
<thead>
<tr>
<th>Inverse Decision Model</th>
<th>Calibration (Low is Better)</th>
<th>PRC Score (High is Better)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black-Box Model:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNN-Based BC-IL</td>
<td>0.18 ± 0.05</td>
<td>0.81 ± 0.08</td>
</tr>
<tr>
<td>IOHMM-Based BC-IL</td>
<td>0.19 ± 0.07</td>
<td>0.79 ± 0.11</td>
</tr>
<tr>
<td>Joint IOHMM-Based BC-IL</td>
<td>0.17 ± 0.05</td>
<td>0.81 ± 0.09</td>
</tr>
<tr>
<td>Reward-Centric Model:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bayesian PO-IRL</td>
<td>0.23 ± 0.01</td>
<td>0.78 ± 0.09</td>
</tr>
<tr>
<td>Joint Bayesian PO-IRL</td>
<td>0.24 ± 0.01</td>
<td>0.79 ± 0.09</td>
</tr>
<tr>
<td>Boundedly Rational Model:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IBRC (with learned $\alpha, \beta, \eta$)</td>
<td>0.16 ± 0.00</td>
<td>0.77 ± 0.01</td>
</tr>
</tbody>
</table>

**Data Selection** From the ADNI data, we first selected out anomalous cases without a cognitive dementia rating test result, which is almost always taken at every visit by every patient. Second, we also truncated patient trajectories at points where a visit is skipped (that is, if the next visit of a patient does not occur immediately after the 6-monthly period following the previous visit). This selection process leaves 1,626 patients out of the original 1,737, and the median number of consecutive visits for each patient is three. In measuring MRI outcomes, the “average” is defined to be within half a standard deviation of the population mean. Note that this is the same pre-processing method employed for ADNI in [22].
Implementation Details of implementation for benchmark algorithms follow the setup in [22], and are reproduced here: 
**RNN-Based BC-IL:** We train an RNN whose inputs are the observed histories $h$ and whose outputs are the predicted probabilities $\hat{p}(u|h)$ of taking action $u$ given the observed history $h$. The network consists of an LSTM unit of size 64 and a fully-connected hidden layer of size 64. The cross-entropy $\mathcal{L} = -\sum_{n=1}^{N} \sum_{t=1}^{T} \sum_{u \in U} \mathbb{I}(u_t = u) \log \hat{p}(u|h)$ is minimized using the Adam optimizer with a learning rate of 0.001 until convergence (that is, when the loss does not improve for 100 consecutive iterations). 

**Bayesian PO-IRL:** The IOHMM parameters are initialized by sampling uniformly at random. Then, they are estimated and fixed using conventional IOHMM methods. The utility $v$ is initialized as $v^0(s, u) = \varepsilon_{s,u}$, where $\varepsilon_{s,u} \sim \mathcal{N}(0, 0.001^2)$. Then, it is estimated via MCMC sampling, during which new candidate samples are generated by adding Gaussian noise with standard deviation 0.001 to the previous sample. To form the final estimate, we average every 10th sample among the second set of 500 samples, ignoring the first 500 samples. To compute optimal $Q$-values, we use an off-the-shelf POMDP solver https://www.pomdp.org/code/index.html.

**Joint Bayesian PO-IRL:** All parameters are initialized exactly the same way as in Bayesian PO-IRL. Then, both the IOHMM parameters and the utility are estimated jointly via MCMC sampling. In order to generate new candidate samples, with equal probabilities we either sample new IOHMM parameters from the posterior (but without changing $v$) or obtain a new $v$ the same way we do in Bayesian PO-IRL (but without changing the IOHMM parameters). A final estimate is formed the same way as in Bayesian PO-IRL. 

**IOHMM-Based BC-IL:** The IOHMM parameters are initialized by sampling them uniformly at random. Then, they are estimated and fixed using conventional IOHMM methods. Given the IOHMM parameters, we parameterize policies using the method of [22], with the policy parameters $\{\mu_u\}_{u \in U}$ (not to be confused with the occupancy measure $\mu$ as defined in the present work) initialized as $\mu^0_u(s) = (1/|S| + \varepsilon_{s,u})/\sum_{s' \in S}(1/|S| + \varepsilon_{s',u'})$, where $\varepsilon_{s,u'} \sim \mathcal{N}(0, 0.001^2)$. Then, they are estimated according solely to the action likelihoods in using the EM algorithm. The expected log-posterior is maximized using the Adam optimizer with learning rate 0.001 until convergence (that is, when the expected log-posterior does not improve for 100 consecutive iterations). 

**Joint IOHMM-Based BC-IL:** This corresponds exactly to the proposed method of [22] itself, which is similar to IOHMM-Based BC-IL except parameters are trained jointly. All parameters are initialized exactly the same way as before; then, the IOHMM parameters and the policy parameters are estimated jointly according to both the action likelihoods and the observation likelihoods simultaneously. The expected log-posterior is again maximized using the Adam optimizer with a learning rate of 0.001 until convergence (non-improvement for 100 consecutive iterations).

C. Proofs of Propositions

**Lemma 1 (Forward Recursion)** Define the forward operator $F_{\pi,\mu} : \Delta(Z) \rightarrow \Delta(Z)$ such that for any given $\mu \in \Delta(Z)$:

$$F_{\pi,\mu}(\mu)(z) := (1 - \gamma)\rho_0(z) + \gamma (M_{\pi,\mu})_z(z)$$

Then the occupancy $\mu_{\pi,\mu}$ is the (unique) fixed point of $F_{\pi,\mu}$.

**Proof.** Start from the definition of $M_{\pi,\mu}$; episodes are restarted on completion ad infinitum, so we can write $\mu_{\pi,\mu}$ as:

$$\mu_{\pi,\mu}(z) := (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t p(z_t | z_0)$$

Then we obtain the result by simple algebraic manipulation:

$$F_{\pi,\mu}(\mu)(z) := (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t (M_{\pi,\mu})_z(z)$$

For uniqueness, we use the usual conditions—that is, that the process induced by the environment and the agent’s policies is ergodic, with a single closed communicating class.

**Lemma 2 (Backward Recursion)** Define the backward operator $B_{\pi,\mu} : \mathbb{R}^Z \rightarrow \mathbb{R}^Z$ such that for any given $\mu \in \mathbb{R}^Z$:

$$B_{\pi,\mu}(\mu)(z) := E_{s \sim p(z)} \left[ v(s, u) + E_{\tau,\omega \sim \sigma(z, u)} \gamma V(z') \right]$$

Then the (dual) optimal $V$ is the (unique) fixed point of $B_{\pi,\mu}$; this is the value function considering knowledge uncertainty:

$$V^{\phi,\sigma}(z) := \sum_{t=0}^{\infty} \gamma^t E_{s \sim p(z)} \left[ u \sim \pi(z) \tau,\omega \sim \sigma(z, u) \sigma(s_{t+1}, u_{t+1}) \pi(z_{t+1}, \omega_{t+1}) \mid z_0 \right]$$

so we can equivalently write targets $J_{\pi,\mu} = E_{z \sim p(z)} V^{\phi,\sigma}(z)$.

Likewise, we can also define the (state-action) value function $Q^{\phi,\sigma}(z, u)$—that is, $Q^{\phi,\sigma}(z, u) := E_{z \sim p(z)} [v(s, u) + E_{\tau,\omega \sim \sigma(z, u)} \gamma V^{\phi,\sigma}(z')]$ given an action.

**Proof.** Start with the Lagrangian, with $V \in \mathbb{R}^Z$:

$$J_{\pi,\mu} := \sum_{z} \mu(z) - \langle V, \mu - \gamma M_{\pi,\mu} \mu \rangle (1 - \gamma) \rho_0$$

$$= E_{z \sim \mu_{\pi,\mu}} v(s, u) - \langle V, \mu - \gamma M_{\pi,\mu} \mu \rangle (1 - \gamma) \rho_0$$

$$= E_{z \sim \mu_{\pi,\mu}} v(s, u) + E_{z \sim \mu_{\pi,\mu}} \gamma V(z')$$

$$- E_{z \sim \mu_{\pi,\mu}} V(z) + \langle V, (1 - \gamma) \rho_0 \rangle$$

(35)
which allows appealing to the contraction mapping theorem.

Then taking the gradient w.r.t. \( \mu \) and setting it to zero yields:

\[
V(z) = \mathbb{E}_{s \sim \pi(z)} [v(s, u) + \mathbb{E}_{\tau \sim \sigma(z, s, u)} \gamma V'(z')] + \langle V, (1 - \gamma) \rho_0 \rangle
\]

For uniqueness, observe as usual that \( \mathbb{B}_{\pi, \rho} \) is \( \gamma \)-contracting:

\[
\|\mathbb{B}_{\pi, \rho} V - \mathbb{B}_{\pi, \rho} V'\|_{\infty} = \max_{z} \mathbb{E} \left[ \gamma V(z') - \gamma V'(z') \right]
\]

which allows appealing to the contraction mapping theorem.

**Proposition 3 (Backward Recursion)** Define the backward operator \( \mathbb{B}_{\pi, \rho} : \mathbb{R}^{2} \rightarrow \mathbb{R}^{2} \) such that for any given function \( V \in \mathbb{R}^{2} \) and for any given coefficient values \( \alpha, \beta, \eta \in \mathbb{R} \):

\[
\mathbb{B}_{\pi, \rho}(V)(z) = \mathbb{E}_{s \sim \pi(z)} \left[ - \alpha \log \frac{\pi(u|z)}{\pi(u)} + v(s, u) + \mathbb{E}_{\tau \sim \sigma(s, u)} \left[ - \beta \log \frac{\sigma(u|z, s)}{\sigma(u)} + \eta \log \frac{\theta_{\tau, \omega}(z'|z, u)}{\theta(z')} + \gamma V'(z') \right] \right]
\]

Then the (dual) optimal \( V \) is the (unique) fixed point of \( \mathbb{B}_{\pi, \rho} \); as before, this is the **value function** \( V^{\pi, \rho} \) which now includes the complexity terms. Likewise, we can also define the (state-action) \( Q^{\pi, \rho} \in \mathbb{R}^{2} \) as the \( 1/3 \)-step-ahead expectation, and the (state-action-model) \( K^{\pi, \rho} \in \mathbb{R}^{2} \times T \times C \) as the \( 2/3 \)-steps-ahead expectation (which is new in this setup).

**Proof.** Start with the Lagrangian, now with the new multipliers \( \alpha, \beta, \eta \in \mathbb{R} \) in addition to \( V \in \mathbb{R}^{2} \): \( \mathcal{L}_{\pi, \rho}(\alpha, \beta, \eta, V) \)

\[
= \mathbb{E} z \sim \mu_{\pi, \rho} \mathbb{E}_{u \sim \pi(z)} [v(s, u) + \mathbb{E}_{\tau \sim \sigma(z, s, u)} \gamma V'(z')] + \langle V, (1 - \gamma) \rho_0 \rangle
\]

Then taking the gradient w.r.t. \( \mu \) and setting it to zero yields:

\[
V(z) = \mathbb{E}_{s \sim \pi(z)} [v(s, u) + \mathbb{E}_{\tau \sim \sigma(z, s, u)} \gamma V'(z')] + \langle V, (1 - \gamma) \rho_0 \rangle
\]

For uniqueness, observe as before that \( \mathbb{B}_{\pi, \rho} \) is \( \gamma \)-contracting:

\[
\|\mathbb{B}_{\pi, \rho} V - \mathbb{B}_{\pi, \rho} V'\|_{\infty} = \max_{z} \mathbb{E} \left[ \gamma V(z') - \gamma V'(z') \right]
\]

which allows appealing to the contraction mapping theorem.

**Theorem 4 (Boundedly Rational Values)** Define the backward operator \( \mathbb{B}^{\ast} : \mathbb{R}^{2} \rightarrow \mathbb{R}^{2} \) such that for any \( V \in \mathbb{R}^{2} \):

\[
\mathbb{B}^{\ast}(V)(z) = \alpha \log \frac{\pi(u|z)}{\pi(u)} + v(s, u) + \mathbb{E}_{\tau \sim \sigma(z, s, u)} \left[ - \beta \log \frac{\sigma(u|z, s)}{\sigma(u)} + \eta \log \frac{\theta_{\tau, \omega}(z'|z, u)}{\theta(z')} + \gamma V'(z') \right]
\]

For uniqueness, observe as before that \( \mathbb{B}_{\pi, \rho} \) is \( \gamma \)-contracting:

\[
\|\mathbb{B}_{\pi, \rho} V - \mathbb{B}_{\pi, \rho} V'\|_{\infty} \leq \gamma \|V - V'\|_{\infty}
\]

then appeal to the contraction mapping theorem for uniqueness of fixed point.

The only change from before is the additional log terms, which—like the utility term—cancel out of the differences.

For Theorems 4 and 5, we give a single derivation for both:

**Theorem 4** (Boundedly Rational Values) Define the backward operator \( \mathbb{B}^{\ast} : \mathbb{R}^{2} \rightarrow \mathbb{R}^{2} \) such that for any \( V \in \mathbb{R}^{2} \):

\[
\mathbb{B}^{\ast}(V)(z) = \alpha \log \frac{\pi(u|z)}{\pi(u)} + v(s, u) + \mathbb{E}_{\tau \sim \sigma(z, s, u)} \left[ - \beta \log \frac{\sigma(u|z, s)}{\sigma(u)} + \eta \log \frac{\theta_{\tau, \omega}(z'|z, u)}{\theta(z')} + \gamma V'(z') \right]
\]
Theorem 5 (Boundedly Rational Policies) The boundedly rational policy (i.e., primal optimal) is given by:
\[ \pi^*(u|z) = \frac{\pi(u)}{Z_{Q^*}(z)} \exp \left( \frac{1}{\alpha} Q^*(z, u) \right) \]  
(21)
and the boundedly rational recognition policy is given by:
\[ \rho^*(z'|z, u, x') = \mathbb{E}_{\tau, \omega \sim \sigma^*(z, u)} \rho_{\tau, \omega}(z'|z, u, x'), \]
where
\[ \sigma^*(\tau, \omega|z, u) = \frac{\sigma(\tau, \omega)}{Z_{K^*}(z, u, \tau, \omega)} \exp \left( \frac{1}{\beta} K^*(z, u, \tau, \omega) \right) \]  
(22)
where \( Z_{Q^*}(z) = \mathbb{E}_{u_{\sim \pi}} \exp \left( \frac{1}{\alpha} Q^*(z, u) \right) \)
and \( Z_{K^*}(z, u) = \mathbb{E}_{\tau, \omega \sim \sigma} \exp \left( \frac{1}{\beta} K^*(z, u, \tau, \omega) \right) \)

which proves Theorem 5. Then Theorem 4 is obtained by plugging back into the backward recursion (Proposition 3).

For uniqueness, we want \( \| B^* V - B^* V' \|_\infty \leq \gamma \| V - V' \|_\infty \). Let \( \| V - V' \|_\infty = \varepsilon \) \( (\max_z | V(z') - V'(z') | = \varepsilon) \). Now, \( (B^* V)(z) \)
\[ \geq \alpha \log \mathbb{E}_{u_{\sim \pi}} \left[ \exp \left( \frac{1}{\alpha} \mathbb{E}_{s \sim p(z|z)} v(s, u) \right) \right] + \beta \log \mathbb{E}_{\tau, \omega \sim \sigma} \left[ \exp \left( \frac{1}{\beta} \mathbb{E}_{s_{\sim \pi} \sim p(z|z)} v(s, u) \right) \right] \]
\[ \leq \alpha \log \mathbb{E}_{u_{\sim \pi}} \left[ \exp \left( \frac{1}{\alpha} \mathbb{E}_{s \sim p(z|z)} v(s, u) \right) \right] + \beta \log \mathbb{E}_{\tau, \omega \sim \sigma} \left[ \exp \left( \frac{1}{\beta} \mathbb{E}_{s_{\sim \pi} \sim p(z|z)} v(s, u) \right) \right] \]
and (state-action-model) \( K^* \in \mathbb{R}^{X \times T \times D} \) by 2/3 steps:
\[ K^{* \pi, \sigma}(z, u, \tau, \omega) \]
\[ \geq \mathbb{E}_{s \sim p(z|z)} \left[ - \eta \log \frac{\theta_{\tau, \omega}(z'|z, u)}{\theta(z')} + \gamma V^{\pi, \sigma}(z') \right] \]
(44)
The decision and recognition policies seek the optimizations:
\[ \text{extremize} \pi V^{\pi, \sigma}(z) \]
\[ \text{s.t.} \quad \mathbb{E}_{u_{\sim \pi} \sim p(z|z)} 1 = 1 \]  
(45)
\[ \text{extremize} \sigma Q^{\pi, \sigma}(z, u) \]
\[ \text{s.t.} \quad \mathbb{E}_{\tau, \omega \sim \sigma} 1 = 1 \]  
(46)
Equations 42-44 are true in particular for optimal values, so
\[ V^*(z) = \mathbb{E}_{u_{\sim \pi} \sim p(z|z)} \left[ - \alpha \log \frac{\pi^*(u|z)}{\pi(u)} + Q^*(z, u) \right] \]  
(47)
\[ Q^*(z, u) = \mathbb{E}_{s \sim p(z|z)} v(s, u) + \mathbb{E}_{\tau, \omega \sim \sigma} \left[ - \beta \log \frac{\sigma^*(\tau, \omega|z, u)}{\sigma(\tau, \omega)} + K^*(z, u, \tau, \omega) \right] \]  
(48)
Therefore for the extremizations we write the Lagrangians
\[ \mathcal{L}(\pi^*, \lambda) = V^*(z) + \lambda \cdot (\mathbb{E}_{u_{\sim \pi} \sim p(z|z)} 1 - 1) \]  
(49)
\[ \mathcal{L}(\sigma^*, \nu) = Q^*(z, u) + \nu \cdot (\mathbb{E}_{\tau, \omega \sim \sigma} 1 - 1) \]  
(50)
Straightforward algebraic manipulation yields the policies:
\[ \pi^*(u|z) = \frac{\pi(u)}{Z_{Q^*}(z)} \exp \left( \frac{1}{\alpha} Q^*(z, u) \right) \]  
(51)
\[ \sigma^*(\tau, \omega|z, u) = \frac{\sigma(\tau, \omega)}{Z_{K^*}(z, u, \tau, \omega)} \exp \left( \frac{1}{\beta} K^*(z, u, \tau, \omega) \right) \]  
(52)
where partition functions \( Z_{Q^*}(z) \) and \( Z_{K^*}(z, u) \) are given by:
\[ Z_{Q^*}(z) = \mathbb{E}_{u_{\sim \pi} \sim p(z|z)} \exp \left( \frac{1}{\alpha} Q^*(z, u) \right) \]  
(53)
\[ Z_{K^*}(z, u) = \mathbb{E}_{\tau, \omega \sim \sigma} \exp \left( \frac{1}{\beta} K^*(z, u, \tau, \omega) \right) \]  
(54)
Likewise, we can show that \( (B^* V)(z) \geq (B^* V')(z) - \gamma \varepsilon \). Hence \( \max_z [(B^* V)(z) - (B^* V')(z)] = (B^* V - B^* V') \leq \gamma \varepsilon \).
Note on Equation 24: Note that we originally formulated “soft policy matching” in Table 3 as a forward Kullback-Leibler divergence expression. However, analogously to maximum likelihood in supervised learning, the entropy terms drop out of the optimization, which yields Equation 24. To see this, note that the causally-conditioned probability is simply the product of conditional probabilities at each time step, and each conditional is “Markovianized” using beliefs $z_t$ (i.e. Equation 25).

D. Illustrative Trajectories

Here we direct attention to the potential utility of IBRC (and—more generally—instantiations of the IDM paradigm) as an “investigative device” for auditing and quantifying individual decisions. In Figure 7, we see that modeling the evolution of a decision-maker’s subjective beliefs provides a concrete basis for analyzing the corresponding sequence of actions chosen. Each vertex of the belief simplex corresponds to one of the three stable Alzheimer’s diagnoses, and each point within the simplex corresponds to a unique belief (i.e. probability distribution). The closer the point is to a vertex (i.e. disease state), the higher the probability assigned to that state. For instance, if the belief is located exactly in the middle of the simplex (i.e. equidistant from all vertices), then all states are believed to be equally likely. Note that this is visual presentation is done similarly to [22], where decision trajectories within belief simplices are first visualized in this manner—with the core difference here being that the decision policies (hence decision boundaries thereby induced) are computed using a different technique.

![Figure 7. Decision Trajectories](image)

Figure 7. Decision Trajectories. Examples of apparent beliefs and actions of a clinical decision-maker regarding real patients, including cases where: (a) the clinician’s decisions coincide with those that would have been dictated by a “perfectly-rational” policy—despite their bounded rationality; (b) the clinician fails to make “perfectly-rational” decisions (in this context, the “boundedness” of the clinician could be due to any number of issues encountered during the diagnostic process); and (c) a patient who—apparently—could have been diagnosed much earlier than they actually were, but for the clinician not having followed the decisions prescribed by the “perfectly-rational” policy.

E. Summary of Notation

<table>
<thead>
<tr>
<th>Notation</th>
<th>Meaning</th>
<th>(first defined in)</th>
<th>Notation</th>
<th>Meaning</th>
<th>(first defined in)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\psi$</td>
<td>problem setting</td>
<td>Section 3.1</td>
<td>$s$</td>
<td>environment state</td>
<td>Section 3.1</td>
</tr>
<tr>
<td>$x$</td>
<td>environment emission</td>
<td>Section 3.1</td>
<td>$z$</td>
<td>agent state, i.e. belief</td>
<td>Section 3.1</td>
</tr>
<tr>
<td>$u$</td>
<td>agent emission, i.e. action</td>
<td>Section 3.1</td>
<td>$\tau_{env}$</td>
<td>environment transition</td>
<td>Section 3.1</td>
</tr>
<tr>
<td>$\tau$</td>
<td>subjective transition</td>
<td>Section 3.1</td>
<td>$\omega_{env}$</td>
<td>environment emission</td>
<td>Section 3.1</td>
</tr>
<tr>
<td>$\omega$</td>
<td>subjective emission</td>
<td>Section 3.1</td>
<td>$\psi$</td>
<td>utility (i.e. reward) function</td>
<td>Section 3.1</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>discount factor</td>
<td>Section 3.1</td>
<td>$\phi$</td>
<td>behavior</td>
<td>Section 3.1</td>
</tr>
<tr>
<td>$\theta_{demos}$</td>
<td>demonstrated behavior</td>
<td>Section 3.2</td>
<td>$\theta_{imit}$</td>
<td>imitation behavior</td>
<td>Section 3.2</td>
</tr>
<tr>
<td>$\theta$</td>
<td>planning parameter</td>
<td>Section 3.1</td>
<td>$\theta_{norm}$</td>
<td>normative parameter</td>
<td>Section 3.2</td>
</tr>
<tr>
<td>$\rho$</td>
<td>descriptive parameter</td>
<td>Section 3.2</td>
<td>$\pi$</td>
<td>decision policy</td>
<td>Section 3.1</td>
</tr>
<tr>
<td>$F$</td>
<td>recognition policy</td>
<td>Section 3.1</td>
<td>$\sigma$</td>
<td>specification policy</td>
<td>Section 4.1</td>
</tr>
<tr>
<td>$\alpha^{-1}$</td>
<td>flexibility coefficient</td>
<td>Section 4.2</td>
<td>$\beta^{-1}$</td>
<td>optimism coefficient</td>
<td>Section 4.2</td>
</tr>
<tr>
<td>$\epsilon^{-1}$</td>
<td>adaptivity coefficient</td>
<td>Section 4.2</td>
<td>$\hat{\pi}$</td>
<td>action prior</td>
<td>Section 4.2</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>model prior</td>
<td>Section 4.2</td>
<td>$\hat{\sigma}$</td>
<td>belief prior</td>
<td>Section 4.2</td>
</tr>
</tbody>
</table>
References


Inverse Decision Modeling


Inverse Decision Modeling


Inverse Decision Modeling


