Appendices

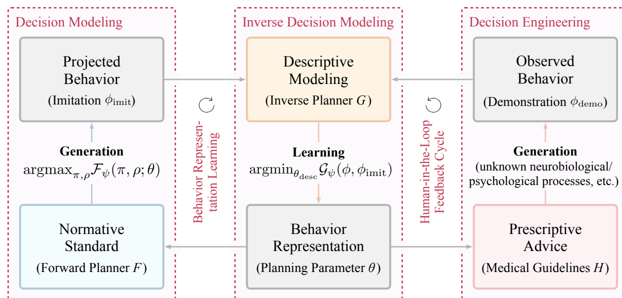
Appendix A gives a longer discussion of merits and caveats; Appendix B gives further experiment details; Appendix C gives derivations of propositions; Appendix D shows illustrative trajectories; Appendix E gives a summary of notation.

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 ϕ_{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 Θ and normative standards θ_{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 ψ); it is also agnostic as to whether the underlying forward problem is model-free or model-based (which—among other properties—is encoded in θ). 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 π as well as a *recognition* policy ρ, 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.

Figure 4. Normative, Prescriptive, and Descriptive Modeling. Recall the "lifecycle" of decision analysis (Section 1). As a paradigm of optimal behavior, *normative standards* serve as a theoretical benchmark. To guide imperfect agents toward this ideal, *prescriptive advice* serves to engineer behavior from humans in the loop. Importantly, however, this first requires an understanding of the imperfections—relative to the normative ideal—that require correcting. This is the goal of *descriptive modeling*—that is, to obtain an empirical account of existing behavior from observed data. Precisely, inverse decision modeling (middle) leverages a normative standard (left) to obtain an interpretable account of demonstrated behavior, thereby enabling the introspection of existing practices, which may inform construction of prescriptive guidelines (right).



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 $\theta_{\text{desc}} = \upsilon$ alone, or assume fully-observable environments (whence $\mathcal{S} = \mathcal{X} = \mathcal{Z}$, and ρ 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. τ, 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 υ 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 Θ (and specification of θ_{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 were parameterized by the (interpretable) structure we designed for Θ. 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 θ_{norm}).

In IBRC, for example, in learning β 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 $\tilde{\sigma}$ to specify as part of θ_{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 $\tilde{\sigma}$ via models learned a priori from data. Of course, coming up such values for θ_{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

Figure 5. Graphical Model. In general, the environment's states (top) are only accessible via its emissions in response to actions (middle), which the agent incorporates by way of internal states (bottom). However, note that—unlike classic POMDP/IOHMM settings, here the agent's knowledge of the dynamics is subjective.

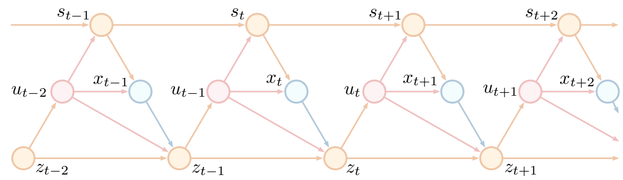
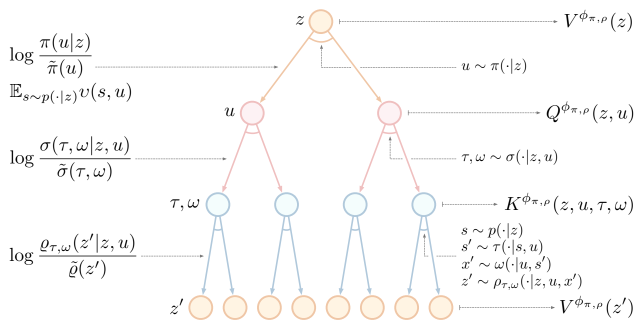


Figure 6. Backup Diagram. In IBRC, the backup operation (Theorem 4) transfers value information across three recursive "layers"—that is, of successor values for agent states (V), state-action pairs (Q), and state-action-model tuples (K). Indicated below are the utility and penalty terms collected along these backup operations.



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 τ, ω). 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 $h \sim \phi_{\pi,\rho}$ 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 υ, 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 υ within θ_{norm}, while recovering one or more behavioral parameters α, β, γ from θ_{desc}. Clearly, however, we are not at liberty to specify completely random values for υ (or, more generally, that we are not at liberty to design Θ and θ_{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 ϕ^*_{imit} onto $\Phi_{\theta_{\text{norm}}}$ may simply lose too much information from ϕ_{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 Θ and θ_{norm}, which requires care in practice. Most importantly, it should be verified that—under our designed parameterization—the *projected* behavior ϕ^*_{imit} is still a faithful model of the *demonstrated* behavior ϕ_{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 (interpretably 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 Φ-space induced by F, G, and Θ, 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 π, to the ideas of subjective dynamics σ and biased belief-updates ρ (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 $\theta_{\text{desc}} = \upsilon$ (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 ζ (cf. Table 3):

$$F^{\text{ME}}_{\theta_{\text{norm}}}(\theta_{\text{desc}}) \doteq \phi_{\pi^*} \text{ where } \pi^* \doteq \operatorname{argmax}_\pi \mathbb{E}_{z\sim\rho_0} V^{\phi_\pi}_{\text{soft},\theta}(z) \quad (26)$$

$$G^{\text{MM}}_{\theta_{\text{norm}}}(\phi) \doteq \operatorname{argmin}_{\theta_{\text{desc}}} \mathbb{E}_{z\sim\rho_0}[V^{\phi_{\text{imit}}}_{\text{soft},\theta}(z) - V^{\phi}_{\theta}(z)] + \zeta(\theta_{\text{desc}}) \quad (27)$$

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

$$F^{\text{DR}}_{\theta_{\text{norm}}}(\theta_{\text{desc}}) \doteq \operatorname{argmax}_\pi \delta(\pi - f_{\text{decision}}(\chi)) \quad (28)$$

$$G^{\text{DM}}_{\theta_{\text{norm}}}(\phi) \doteq \operatorname{argmin}_{\theta_{\text{desc}}} \zeta^*(\phi_{\text{demo}} - \phi_{\text{imit}}) - \mathcal{H}_{\text{imit}} \quad (29)$$

where distance measures are given by the convex conjugate ζ^*, and $\mathcal{H}_{\text{imit}}$ gives the causal entropy of the imitating policy. Now, the primary motivation behind generative adversarial imitation learning is the observation that ζ-regularized maximum-margin soft IRL implicitly seeks a policy whose occupancy is close to the demonstrator's as measured by ζ^*. In IDM, this corresponds to a remarkably simple statement:

[6] Abstractly, this is not dissimilar to any type of model fitting problem: If the mismatch between the (unknown) data generating process and the (imposed) structure of the model is too great, then the quality of the model—by any reasonable measure—would suffer.

Proposition 6 (Ho and Ermon, Recast) Define the *behavior projections* induced by the composition of each pairing:

$$\text{proj}^{\text{ME,MM}}_{\Phi_{\theta_\text{norm}}} \doteq F^{\text{ME}}_{\theta_\text{norm}} \circ G^{\text{MM}}_{\theta_\text{norm}} \tag{30}$$

$$\text{proj}^{\text{DR,DM}}_{\Phi_{\theta_\text{norm}}} \doteq F^{\text{DR}}_{\theta_\text{norm}} \circ G^{\text{DM}}_{\theta_\text{norm}} \tag{31}$$

Then these projections are identical: $\text{proj}^{\text{ME,MM}}_{\Phi_{\theta_\text{norm}}} = \text{proj}^{\text{DR,DM}}_{\Phi_{\theta_\text{norm}}}$ (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 $\theta_\text{desc} = \upsilon$, 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 threefold. 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 interpretably 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 Θ 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 $\mathcal{Z} \doteq \mathbb{R}^k$, 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 Θ_{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 τ, ω, update $\rho_{\tau,\omega}(z'|z,u,x')$ is the Dirac delta centered at

$$p(s'|z,u,x',\tau,\omega) \doteq \mathbb{E}_{s\sim p(\cdot|z)}\left[\frac{\tau(s'|s,u)\omega(x'|u,s')}{\mathbb{E}_{s'\sim\tau(\cdot|s,u)}\omega(x'|u,s')}\right] \quad (32)$$

and the overall recognition policy is the expectation over such values of τ, ω (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 $\pm 10\%$ 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 ϕ^*_{imit} is still an *accurate* model of the demonstrated behavior ϕ_{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 $\tilde{\sigma}$).

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).

Inverse Decision Model	**Calibration (Low is Better)**	**PRC Score (High is Better)**
Black-Box Model:		
RNN-Based BC-IL	0.18 ± 0.05	0.81 ± 0.08
IOHMM-Based BC-IL	0.19 ± 0.07	0.79 ± 0.11
Joint IOHMM-Based BC-IL	0.17 ± 0.05	0.81 ± 0.09
Reward-Centric Model:		
Bayesian PO-IRL	0.23 ± 0.01	0.78 ± 0.09
Joint Bayesian PO-IRL	0.24 ± 0.01	0.79 ± 0.09
Boundedly Rational Model:		
IBRC (with learned α, β, η)	0.16 ± 0.00	0.77 ± 0.01

<u>*Results*</u>. 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), which—for our ADNI example—affirms the validity of IBRC as an (interpretable) representation of ϕ_{demo}.

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 $\hat{\pi}(u|h)$ 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 $\mathcal{L}=-\sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{u\in\mathcal{U}}\mathbb{I}\{u_t=u\}\log\hat{\pi}(u|h)$ 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 υ is initialized as $\hat{\upsilon}^0(s,u)=\varepsilon_{s,u}$, 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 υ) or obtain a new υ 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 $\{\mu_u\}_{u\in\mathcal{U}}$ (not to be confused with the occupancy measure "μ" as defined in the present work) initialized as $\hat{\mu}_u^0(s)=(1/|S|+\varepsilon_{u,s})/\sum_{s'\in S}(1/|S|+\varepsilon_{u,s'})$, where $\varepsilon_{u,s'}\sim\mathcal{N}(0,0.001^2)$. 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 $\mathbb{F}_{\pi,\rho}:\Delta(\mathcal{Z})^{\Delta(\mathcal{Z})}$ such that for any given $\mu\in\Delta(\mathcal{Z})$:

$$(\mathbb{F}_{\pi,\rho}\mu)(z)\doteq(1-\gamma)\rho_0(z)+\gamma(\mathbb{M}_{\pi,\rho}\mu)(z) \tag{12}$$

Then the occupancy $\mu_{\pi,\rho}$ is the (unique) fixed point of $\mathbb{F}_{\pi,\rho}$.

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

$$\begin{aligned}\mu_{\pi,\rho}(z)&\doteq(1-\gamma)\textstyle\sum_{t=0}^{\infty}\gamma^t p(z_t=z|z_0\sim\rho_0)\\&=(1-\gamma)\textstyle\sum_{t=0}^{\infty}\gamma^t((\mathbb{M}_{\pi,\rho})^t\rho_0)(z)\end{aligned} \tag{33}$$

Then we obtain the result by simple algebraic manipulation:

$$\begin{aligned}&(1-\gamma)\rho_0(z)+\gamma(\mathbb{M}_{\pi,\rho}\mu_{\pi,\rho})(z)\\&=(1-\gamma)\rho_0(z)+\gamma(1-\gamma)\textstyle\sum_{t=0}^{\infty}\gamma^t((\mathbb{M}_{\pi,\rho})^{t+1}\rho_0)(z)\\&=(1-\gamma)(\rho_0(z)+\textstyle\sum_{t=0}^{\infty}\gamma^{t+1}((\mathbb{M}_{\pi,\rho})^{t+1}\rho_0)(z))\\&=(1-\gamma)\textstyle\sum_{t=0}^{\infty}\gamma^t((\mathbb{M}_{\pi,\rho})^t\rho_0)(z)\\&=\mu_{\pi,\rho}(z)\end{aligned} \tag{34}$$

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 $\mathbb{B}_{\pi,\rho}:\mathbb{R}^{\mathcal{Z}}\to\mathbb{R}^{\mathcal{Z}}$ such that for any given $V\in\mathbb{R}^{\mathcal{Z}}$:

$$(\mathbb{B}_{\pi,\rho}V)(z)\doteq\mathbb{E}_{\substack{s\sim p(\cdot|z)\\u\sim\pi(\cdot|z)}}[\upsilon(s,u)+\mathbb{E}_{\substack{\tau,\omega\sim\sigma(\cdot|z,u)\\s'\sim\tau(\cdot|s,u)\\x'\sim\omega(\cdot|u,s')\\z'\sim\rho_{\tau,\omega}(\cdot|z,u,x')}}\gamma V(z')] \tag{14}$$

Then the (dual) optimal V is the (unique) fixed point of $\mathbb{B}_{\pi,\rho}$; this is the *value function* considering knowledge uncertainty:

$$V^{\phi_{\pi,\rho}}(z)\doteq\textstyle\sum_{t=0}^{\infty}\gamma^t\mathbb{E}_{\substack{s_t\sim p(\cdot|z_t)\\u_t\sim\pi(\cdot|z_t)\\\tau,\omega\sim\sigma(\cdot|z_t,u_t)\\s_{t+1}\sim\tau(\cdot|s_t,u_t)\\x_{t+1}\sim\omega(\cdot|u_t,s_{t+1})\\z_{t+1}\sim\rho_{\tau,\omega}(\cdot|z_t,u_t,x_{t+1})}}[\upsilon(s_t,u_t)|z_0=z] \tag{15}$$

so we can equivalently write targets $J_{\pi,\rho}=\mathbb{E}_{z\sim\rho_0}V^{\phi_{\pi,\rho}}(z)$. Likewise, we can also define the (state-action) value function $Q^{\phi_{\pi,\rho}}\in\mathbb{R}^{\mathcal{Z}\times\mathcal{U}}$—that is, $Q^{\phi_{\pi,\rho}}(z,u)\doteq\mathbb{E}_{s\sim p(\cdot|z)}[\upsilon(s,u)+\mathbb{E}_{\tau,\omega\sim\sigma(\cdot|z,u),\ldots,z'\sim\rho_{\tau,\omega}(\cdot|z,u,x')}\gamma V^{\phi_{\pi,\rho}}(z')]$ given an action.

Proof. Start with the Lagrangian, with $V\in\mathbb{R}^{\mathcal{Z}}$: $\mathcal{L}_{\pi,\rho}(\mu,V)$

$$\begin{aligned}&\doteq J_{\pi,\rho}-\langle V,\mu-\gamma\mathbb{M}_{\pi,\rho}\mu-(1-\gamma)\rho_0\rangle\\&=\mathbb{E}_{\substack{z\sim\mu_{\pi,\rho}\\s\sim p(\cdot|z)\\u\sim\pi(\cdot|z)}}\upsilon(s,u)-\langle V,\mu-\gamma\mathbb{M}_{\pi,\rho}\mu-(1-\gamma)\rho_0\rangle\\&=\mathbb{E}_{\substack{z\sim\mu_{\pi,\rho}\\s\sim p(\cdot|z)\\u\sim\pi(\cdot|z)}}\upsilon(s,u)+\mathbb{E}_{\substack{z\sim\mu_{\pi,\rho}\\s\sim p(\cdot|z)\\u\sim\pi(\cdot|z)\\\tau,\omega\sim\sigma(\cdot|z,u)\\s'\sim\tau(\cdot|s,u)\\x'\sim\omega(\cdot|u,s')\\z'\sim\rho(\cdot|z,u,x')}}\gamma V(z')\\&\quad-\mathbb{E}_{z\sim\mu_{\pi,\rho}}V(z)+\langle V,(1-\gamma)\rho_0\rangle\end{aligned} \tag{35}$$

$$= \mathbb{E}_{\substack{z\sim\mu_{\pi,\rho}\\ s\sim p(\cdot|z)}}[\mathbb{E}_{u\sim\pi(\cdot|z)}[\upsilon(s,u) \tag{36}$$
$$+ \mathbb{E}_{\substack{\tau,\omega\sim\sigma(\cdot|z,u)\\ s'\sim\tau(\cdot|s,u)\\ x'\sim\omega(\cdot|u,s')\\ z'\sim\rho(\cdot|z,u,x')}} \gamma V(z')] - V(z)] + \langle V,(1-\gamma)\rho_0\rangle$$

Then taking the gradient w.r.t. μ and setting it to zero yields:

$$V(z) = \mathbb{E}_{\substack{s\sim p(\cdot|z)\\ u\sim\pi(\cdot|z)}}[\upsilon(s,u) + \mathbb{E}_{\substack{\tau,\omega\sim\sigma(\cdot|z,u)\\ s'\sim\tau(\cdot|s,u)\\ x'\sim\omega(\cdot|u,s')\\ z'\sim\rho(\cdot|z,u,x')}} \gamma V(z')] \tag{37}$$

For uniqueness, observe as usual that $\mathbb{B}_{\pi,\rho}$ is γ-contracting:

$$\begin{aligned}
&\|\mathbb{B}_{\pi,\rho}V - \mathbb{B}_{\pi,\rho}V'\|_\infty \\
&\quad= \max_z \big|\mathbb{E}_{\substack{u\sim\pi(\cdot|z)\\ \tau,\omega\sim\sigma(\cdot|z,u)\\ z'\sim\varrho_{\tau,\omega}(\cdot|z,u)}} \big[\gamma V(z') - \gamma V'(z')\big]\big| \\
&\quad\leq \max_z \mathbb{E}_{\substack{u\sim\pi(\cdot|z)\\ \tau,\omega\sim\sigma(\cdot|z,u)\\ z'\sim\varrho_{\tau,\omega}(\cdot|z,u)}} \big[\big|\gamma V(z') - \gamma V'(z')\big|\big] \qquad (38)\\
&\quad\leq \max_{z'} \big|\gamma V(z') - \gamma V'(z')\big| \\
&\quad= \gamma\|V - V'\|_\infty
\end{aligned}$$

which allows appealing to the contraction mapping theorem.

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

$$\begin{aligned}
(\mathbb{B}_{\pi,\rho}V)(z) \doteq\ & \mathbb{E}_{\substack{s\sim p(\cdot|z)\\ u\sim\pi(\cdot|z)}}\big[-\alpha\log\frac{\pi(u|z)}{\tilde{\pi}(u)} + \upsilon(s,u) + \\
& \mathbb{E}_{\tau,\omega\sim\sigma(\cdot|z,u)}\big[-\beta\log\frac{\sigma(\tau,\omega|z,u)}{\tilde{\sigma}(\tau,\omega)} + \qquad (19)\\
& \mathbb{E}_{\substack{s'\sim\tau(\cdot|s,u)\\ x'\sim\omega(\cdot|u,s')\\ z'\sim\rho_{\tau,\omega}(\cdot|z,u,x')}} \big[-\eta\log\frac{\varrho_{\tau,\omega}(z'|z,u)}{\tilde{\varrho}(z')} + \gamma V(z')\big]\big]\big]
\end{aligned}$$

Then the (dual) optimal V is the (unique) fixed point of $\mathbb{B}_{\pi,\rho}$; as before, this is the *value function* $V^{\phi_{\pi,\rho}}$—which now includes the complexity terms. Likewise, we can also define the (state-action) $Q^{\phi_{\pi,\rho}} \in \mathbb{R}^{\mathcal{Z}\times\mathcal{U}}$ as the 1/3-step-ahead expectation, and the (state-action-model) $K^{\phi_{\pi,\rho}} \in \mathbb{R}^{\mathcal{Z}\times\mathcal{U}\times\mathcal{T}\times\mathcal{O}}$ 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 $V \in \mathbb{R}^\mathcal{Z}$: $\mathcal{L}_{\pi,\rho}(\mu,\alpha,\beta,\eta,V)$

$$\begin{aligned}
&\doteq J_{\pi,\rho} - \langle V, \mu - \gamma\mathbb{M}_{\pi,\rho}\mu - (1-\gamma)\rho_0\rangle \\
&\quad - \alpha\cdot(\mathbb{I}_{\pi,\rho}[\pi;\tilde{\pi}] - A) - \beta\cdot(\mathbb{I}_{\pi,\rho}[\sigma;\tilde{\sigma}] - B) \\
&\quad - \eta\cdot(\mathbb{I}_{\pi,\rho}[\varrho;\tilde{\varrho}] - C) \\
&= \mathbb{E}_{\substack{z\sim\mu_{\pi,\rho}\\ s\sim p(\cdot|z)\\ u\sim\pi(\cdot|z)}} \upsilon(s,u) - \langle V, \mu - \gamma\mathbb{M}_{\pi,\rho}\mu - (1-\gamma)\rho_0\rangle \\
&\quad - \alpha\cdot(\mathbb{E}_{z\sim\mu_{\pi,\rho}} D_{\mathrm{KL}}(\pi(\cdot|z)\|\tilde{\pi}) - A) \\
&\quad - \beta\cdot(\mathbb{E}_{\substack{z\sim\mu_{\pi,\rho}\\ u\sim\pi(\cdot|z)}} D_{\mathrm{KL}}(\sigma(\cdot|z,u)\|\tilde{\sigma}) - B) \\
&\quad - \eta\cdot(\mathbb{E}_{\substack{z\sim\mu_{\pi,\rho}\\ u\sim\pi(\cdot|z)\\ \tau,\omega\sim\sigma(\cdot|z,u)}} D_{\mathrm{KL}}(\varrho_{\tau,\omega}(\cdot|z,u)\|\tilde{\varrho}) - C) \qquad (39)
\end{aligned}$$

$$\begin{aligned}
&= \mathbb{E}_{\substack{z\sim\mu_{\pi,\rho}\\ s\sim p(\cdot|z)\\ u\sim\pi(\cdot|z)}} \upsilon(s,u) + \mathbb{E}_{\substack{z\sim\mu_{\pi,\rho}\\ s\sim p(\cdot|z)\\ u\sim\pi(\cdot|z)\\ \tau,\omega\sim\sigma(\cdot|z,u)\\ s'\sim\tau(\cdot|s,u)\\ x'\sim\omega(\cdot|u,s')\\ z'\sim\rho(\cdot|z,u,x')}} \gamma V(z') \\
&\quad - \mathbb{E}_{z\sim\mu_{\pi,\rho}} V(z) + \langle V,(1-\gamma)\rho_0\rangle \\
&\quad - \alpha\cdot(\mathbb{E}_{\substack{z\sim\mu_{\pi,\rho}\\ u\sim\pi(\cdot|z)}} \log\frac{\pi(u|z)}{\tilde{\pi}(u)} - A) \\
&\quad - \beta\cdot(\mathbb{E}_{\substack{z\sim\mu_{\pi,\rho}\\ u\sim\pi(\cdot|z)\\ \tau,\omega\sim\sigma(\cdot|z,u)}} \log\frac{\sigma(\tau,\omega|z,u)}{\tilde{\sigma}(\tau,\omega)} - B) \\
&\quad - \eta\cdot(\mathbb{E}_{\substack{z\sim\mu_{\pi,\rho}\\ u\sim\pi(\cdot|z)\\ \tau,\omega\sim\sigma(\cdot|z,u)\\ s'\sim\tau(\cdot|s,u)\\ x'\sim\omega(\cdot|u,s')\\ z'\sim\rho(\cdot|z,u,x')}} \log\frac{\varrho_{\tau,\omega}(z'|z,u)}{\tilde{\varrho}(z')} - C) \\
&= \mathbb{E}_{\substack{z\sim\mu_{\pi,\rho}\\ s\sim p(\cdot|z)}}\big[\mathbb{E}_{u\sim\pi(\cdot|z)}\big[\upsilon(s,u) - \alpha\cdot(\log\frac{\pi(u|z)}{\tilde{\pi}(u)} - A) \\
&\quad + \mathbb{E}_{\tau,\omega\sim\sigma(\cdot|z,u)}\big[-\beta\cdot(\log\frac{\sigma(\tau,\omega|z,u)}{\tilde{\sigma}(\tau,\omega)} - B) \\
&\quad + \mathbb{E}_{\substack{s'\sim\tau(\cdot|s,u)\\ x'\sim\omega(\cdot|u,s')\\ z'\sim\rho(\cdot|z,u,x')}} \big[-\eta\cdot(\log\frac{\varrho_{\tau,\omega}(z'|z,u)}{\tilde{\varrho}(z')} - C) \\
&\quad + \gamma V(z')\big]\big]\big] - V(z)\big] + \langle V,(1-\gamma)\rho_0\rangle \qquad (40)
\end{aligned}$$

Then taking the gradient w.r.t. μ and setting it to zero yields:

$$\begin{aligned}
V(z) =\ & \mathbb{E}_{\substack{s\sim p(\cdot|z)\\ u\sim\pi(\cdot|z)}}\big[-\alpha\log\frac{\pi(u|z)}{\tilde{\pi}(u)} + \upsilon(s,u) + \\
& \mathbb{E}_{\tau,\omega\sim\sigma(\cdot|z,u)}\big[-\beta\log\frac{\sigma(\tau,\omega|z,u)}{\tilde{\sigma}(\tau,\omega)} + \qquad (41)\\
& \mathbb{E}_{\substack{s'\sim\tau(\cdot|s,u)\\ x'\sim\omega(\cdot|u,s')\\ z'\sim\rho_{\tau,\omega}(\cdot|z,u,x')}} \big[-\eta\log\frac{\varrho_{\tau,\omega}(z'|z,u)}{\tilde{\varrho}(z')} + \gamma V(z')\big]\big]\big]
\end{aligned}$$

For uniqueness, observe as before that $\mathbb{B}_{\pi,\rho}$ is γ-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 $\mathbb{B}^*: \mathbb{R}^\mathcal{Z} \to \mathbb{R}^\mathcal{Z}$ such that for any $V \in \mathbb{R}^\mathcal{Z}$:

$$\begin{aligned}
(\mathbb{B}^*V)(z) &\doteq \alpha\log\mathbb{E}_{u\sim\tilde{\pi}}\exp(\tfrac{1}{\alpha}Q(z,u)) \qquad (20)\\
Q(z,u) &\doteq \beta\log\mathbb{E}_{\tau,\omega\sim\tilde{\sigma}}\exp(\tfrac{1}{\beta}K(z,u,\tau,\omega)) \\
K(z,u,\tau,\omega) &\doteq + \mathbb{E}_{s\sim p(\cdot|z)}\upsilon(s,u) \\
& \mathbb{E}_{\substack{s\sim p(\cdot|z)\\ s'\sim\tau(\cdot|s,u)\\ x'\sim\omega(\cdot|u,s')\\ z'\sim\rho_{\tau,\omega}(\cdot|z,u,x')}} \big[-\eta\log\frac{\varrho_{\tau,\omega}(z'|z,u)}{\tilde{\varrho}(z')} + \gamma V(z')\big]
\end{aligned}$$

Then the *boundedly rational value function* V^* for the (primal) optimal π^*,ρ^* is the (unique) fixed point of $\mathbb{B}^*_{\pi,\rho}$. (Note that both Q^* and K^* are immediately obtainable from this).

Theorem 5 (Boundedly Rational Policies) The *boundedly rational decision policy* (i.e. primal optimal) is given by:

$$\pi^*(u|z) = \frac{\tilde{\pi}(u)}{Z_{Q^*}(z)} \exp\big(\tfrac{1}{\alpha}Q^*(z,u)\big) \tag{21}$$

and the *boundedly rational recognition policy* is given by:

$$\rho^*(z'|z,u,x') = \mathbb{E}_{\tau,\omega\sim\sigma^*(\cdot|z,u)}\rho_{\tau,\omega}(z'|z,u,x')\text{ , where}$$
$$\sigma^*(\tau,\omega|z,u) \doteq \frac{\tilde{\sigma}(\tau,\omega)}{Z_{K^*}(z,u)} \exp\big(\tfrac{1}{\beta}K^*(z,u,\tau,\omega)\big) \tag{22}$$

where $Z_{Q^*}(z) = \mathbb{E}_{u\sim\tilde{\pi}}\exp(\frac{1}{\alpha}Q^*(z,u))$ and $Z_{K^*}(z,u) = \mathbb{E}_{\tau,\omega\sim\tilde{\sigma}}\exp(\frac{1}{\beta}K^*(z,u,\tau,\omega))$ give the partition functions.

Proof. From Proposition 3, the (state) value $V^{\phi_{\pi,\rho}} \in \mathbb{R}^{\mathcal{Z}}$ is:

$$\begin{aligned} V^{\phi_{\pi,\rho}}(z) = \ & \mathbb{E}_{\substack{s\sim p(\cdot|z)\\ u\sim\pi(\cdot|z)}}\big[-\alpha\log\frac{\pi(u|z)}{\tilde{\pi}(u)} + \upsilon(s,u) + \\ & \mathbb{E}_{\tau,\omega\sim\sigma(\cdot|z,u)}\big[-\beta\log\frac{\sigma(\tau,\omega|z,u)}{\tilde{\sigma}(\tau,\omega)} + \\ & \mathbb{E}_{\substack{s'\sim\tau(\cdot|s,u)\\ x'\sim\omega(\cdot|u,s')\\ z'\sim\rho_{\tau,\omega}(\cdot|z,u,x')}}\big[-\eta\log\frac{\varrho_{\tau,\omega}(z'|z,u)}{\tilde{\varrho}(z')} + \gamma V^{\phi_{\pi,\rho}}(z')\big]\big]\big] \end{aligned} \tag{42}$$

Define (state-action) $Q^{\phi_{\pi,\rho}} \in \mathbb{R}^{\mathcal{Z}\times\mathcal{U}}$ to be ahead by 1/3 steps:

$$\begin{aligned} Q^{\phi_{\pi,\rho}}(z,u) \doteq \ & \mathbb{E}_{s\sim p(\cdot|z)}\big[\upsilon(s,u) + \\ & \mathbb{E}_{\tau,\omega\sim\sigma(\cdot|z,u)}\big[-\beta\log\frac{\sigma(\tau,\omega|z,u)}{\tilde{\sigma}(\tau,\omega)} + \\ & \mathbb{E}_{\substack{s'\sim\tau(\cdot|s,u)\\ x'\sim\omega(\cdot|u,s')\\ z'\sim\rho_{\tau,\omega}(\cdot|z,u,x')}}\big[-\eta\log\frac{\varrho_{\tau,\omega}(z'|z,u)}{\tilde{\varrho}(z')} + \gamma V^{\phi_{\pi,\rho}}(z')\big]\big]\big] \end{aligned} \tag{43}$$

and (state-action-model) $K^{\phi_{\pi,\rho}} \in \mathbb{R}^{\mathcal{Z}\times\mathcal{U}\times\mathcal{T}\times\mathcal{O}}$ by 2/3 steps:

$$\begin{aligned} & K^{\phi_{\pi,\rho}}(z,u,\tau,\omega) \doteq \\ & \mathbb{E}_{\substack{s\sim p(\cdot|z)\\ s'\sim\tau(\cdot|s,u)\\ x'\sim\omega(\cdot|u,s')\\ z'\sim\rho_{\tau,\omega}(\cdot|z,u,x')}}\big[-\eta\log\frac{\varrho_{\tau,\omega}(z'|z,u)}{\tilde{\varrho}(z')} + \gamma V^{\phi_{\pi,\rho}}(z')\big]\big]\big] \end{aligned} \tag{44}$$

The decision and recognition policies seek the optimizations:

$$\begin{aligned} &\text{extremize}_\pi V^{\phi_{\pi,\rho}}(z) \\ &\text{s.t.} \quad \mathbb{E}_{u\sim\pi(\cdot|z)}1 = 1 \end{aligned} \tag{45}$$

$$\begin{aligned} &\text{extremize}_\sigma Q^{\phi_{\pi,\rho}}(z,u) \\ &\text{s.t.} \quad \mathbb{E}_{\tau,\omega\sim\sigma(\cdot|z,u)}1 = 1 \end{aligned} \tag{46}$$

Equations 42–44 are true in particular for optimal values, so

$$V^*(z) = \mathbb{E}_{u\sim\pi^*(\cdot|z)}\big[-\alpha\log\frac{\pi^*(u|z)}{\tilde{\pi}(u)} + Q^*(z,u)\big] \tag{47}$$

$$\begin{aligned} Q^*(z,u) = \ & \mathbb{E}_{s\sim p(\cdot|z)}\big[\upsilon(s,u)\big] + \mathbb{E}_{\tau,\omega\sim\sigma^*(\cdot|z,u)}\big[\\ & -\beta\log\frac{\sigma^*(\tau,\omega|z,u)}{\tilde{\sigma}(\tau,\omega)} + K^*(z,u,\tau,\omega)\big] \end{aligned} \tag{48}$$

Therefore for the extremizations we write the Lagrangians

$$\mathcal{L}(\pi^*,\lambda) \doteq V^*(z) + \lambda\cdot(\mathbb{E}_{u\sim\pi^*(\cdot|z)}1 - 1) \tag{49}$$

$$\mathcal{L}(\sigma^*,\nu) \doteq Q^*(z,u) + \nu\cdot(\mathbb{E}_{\tau,\omega\sim\sigma^*(\cdot|z,u)}1 - 1) \tag{50}$$

Straightforward algebraic manipulation yields the policies:

$$\pi^*(u|z) = \frac{\tilde{\pi}(u_t)}{Z_{Q^*}(z)} \exp\big(\tfrac{1}{\alpha}Q^*(z,u)\big) \tag{51}$$

$$\sigma^*(\tau,\omega|z,u) = \frac{\tilde{\sigma}(\tau,\omega)}{Z_{K^*}(z,u)} \exp\big(\tfrac{1}{\beta}K^*(z,u,\tau,\omega)\big) \tag{52}$$

where partition functions $Z_{Q^*}(z)$ and $Z_{K^*}(z)$ are given by:

$$Z_{Q^*}(z) = \mathbb{E}_{u\sim\tilde{\pi}}\exp(\tfrac{1}{\alpha}Q^*(z,u)) \tag{53}$$

$$Z_{K^*}(z,u) = \mathbb{E}_{\tau,\omega\sim\tilde{\sigma}}\exp(\tfrac{1}{\beta}K^*(z,u,\tau,\omega)) \tag{54}$$

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

For uniqueness, we want $\|\mathbb{B}V - \mathbb{B}V'\|_\infty \leq \gamma\|V-V'\|_\infty$. Let $\|V-V'\|_\infty = \varepsilon$ ($\max_{z'}|V(z')-V'(z')| = \varepsilon$). Now, $(\mathbb{B}^*V)(z)$

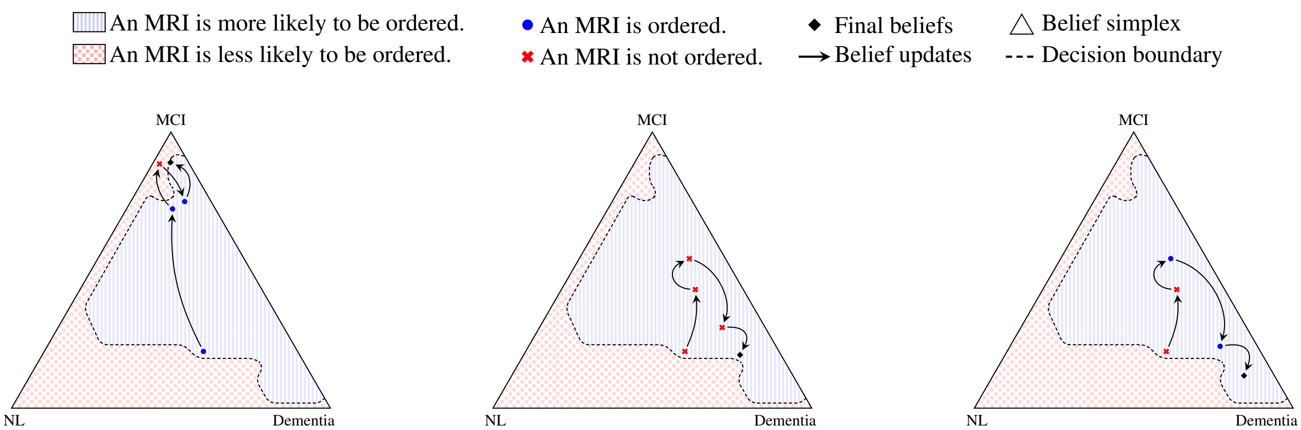
$$\begin{aligned}
&\doteq \alpha\log\mathbb{E}_{u\sim\tilde{\pi}}\big[\exp\big(\tfrac{1}{\alpha}\big(\mathbb{E}_{s\sim p(\cdot|z)}\upsilon(s,u) \\
&\quad + \beta\log\mathbb{E}_{\tau,\omega\sim\tilde{\sigma}}\big[\exp\big(\tfrac{1}{\beta}\mathbb{E}_{z'\sim\varrho_{\tau,\omega}(\cdot|z,u)}\big[\\
&\quad - \eta\log\frac{\varrho_{\tau,\omega}(z'|z,u)}{\tilde{\varrho}(z')} + \gamma V(z')\big]\big)\big]\big)\big)\big] \\
&\leq \alpha\log\mathbb{E}_{u\sim\tilde{\pi}}\big[\exp\big(\tfrac{1}{\alpha}\big(\mathbb{E}_{s\sim p(\cdot|z)}\upsilon(s,u) \\
&\quad + \beta\log\mathbb{E}_{\tau,\omega\sim\tilde{\sigma}}\big[\exp\big(\tfrac{1}{\beta}\mathbb{E}_{z'\sim\varrho_{\tau,\omega}(\cdot|z,u)}\big[\\
&\quad - \eta\log\frac{\varrho_{\tau,\omega}(z'|z,u)}{\tilde{\varrho}(z')} + \gamma(V'(z')+\varepsilon)\big]\big)\big]\big)\big)\big] \\
&= \alpha\log\mathbb{E}_{u\sim\tilde{\pi}}\big[\exp\big(\tfrac{1}{\alpha}\big(\mathbb{E}_{s\sim p(\cdot|z)}\upsilon(s,u) \\
&\quad + \beta\log\mathbb{E}_{\tau,\omega\sim\tilde{\sigma}}\big[\exp\big(\tfrac{1}{\beta}\gamma\varepsilon + \tfrac{1}{\beta}\mathbb{E}_{z'\sim\varrho_{\tau,\omega}(\cdot|z,u)}\big[\\
&\quad - \eta\log\frac{\varrho_{\tau,\omega}(z'|z,u)}{\tilde{\varrho}(z')} + \gamma V'(z')\big]\big)\big]\big)\big)\big] \\
&= \alpha\log\mathbb{E}_{u\sim\tilde{\pi}}\big[\exp\big(\tfrac{1}{\alpha}\big(\mathbb{E}_{s\sim p(\cdot|z)}\upsilon(s,u) \\
&\quad + \beta\log\big(\exp(\tfrac{1}{\beta}\gamma\varepsilon)\mathbb{E}_{\tau,\omega\sim\tilde{\sigma}}\big[\exp\big(\tfrac{1}{\beta}\mathbb{E}_{z'\sim\varrho_{\tau,\omega}(\cdot|z,u)}\big[\\
&\quad - \eta\log\frac{\varrho_{\tau,\omega}(z'|z,u)}{\tilde{\varrho}(z')} + \gamma V'(z')\big]\big)\big]\big)\big)\big)\big] \\
&= \alpha\log\mathbb{E}_{u\sim\tilde{\pi}}\big[\exp\big(\tfrac{1}{\alpha}\gamma\varepsilon + \tfrac{1}{\alpha}\big(\mathbb{E}_{s\sim p(\cdot|z)}\upsilon(s,u) \\
&\quad + \beta\log\mathbb{E}_{\tau,\omega\sim\tilde{\sigma}}\big[\exp\big(\tfrac{1}{\beta}\mathbb{E}_{z'\sim\varrho_{\tau,\omega}(\cdot|z,u)}\big[\\
&\quad - \eta\log\frac{\varrho_{\tau,\omega}(z'|z,u)}{\tilde{\varrho}(z')} + \gamma V'(z')\big]\big)\big]\big)\big)\big] \\
&= \alpha\log\big(\exp(\tfrac{1}{\alpha}\gamma\varepsilon)\mathbb{E}_{u\sim\tilde{\pi}}\big[\exp\big(\tfrac{1}{\alpha}\big(\mathbb{E}_{s\sim p(\cdot|z)}\upsilon(s,u) \\
&\quad + \beta\log\mathbb{E}_{\tau,\omega\sim\tilde{\sigma}}\big[\exp\big(\tfrac{1}{\beta}\mathbb{E}_{z'\sim\varrho_{\tau,\omega}(\cdot|z,u)}\big[\\
&\quad - \eta\log\frac{\varrho_{\tau,\omega}(z'|z,u)}{\tilde{\varrho}(z')} + \gamma V'(z')\big]\big)\big]\big)\big)\big]\big) \\
&= \gamma\varepsilon + \alpha\log\mathbb{E}_{u\sim\tilde{\pi}}\big[\exp\big(\tfrac{1}{\alpha}\big(\mathbb{E}_{s\sim p(\cdot|z)}\upsilon(s,u) \\
&\quad + \beta\log\mathbb{E}_{\tau,\omega\sim\tilde{\sigma}}\big[\exp\big(\tfrac{1}{\beta}\mathbb{E}_{z'\sim\varrho_{\tau,\omega}(\cdot|z,u)}\big[\\
&\quad - \eta\log\frac{\varrho_{\tau,\omega}(z'|z,u)}{\tilde{\varrho}(z')} + \gamma V'(z')\big]\big)\big]\big)\big)\big] \\
&= \gamma\varepsilon + (\mathbb{B}^*V')(z)
\end{aligned} \tag{55}$$

Likewise, we can show that $(\mathbb{B}^*V)(z) \geq (\mathbb{B}^*V')(z) - \gamma\varepsilon$. Hence $\max_z|(\mathbb{B}V)(z) - (\mathbb{B}V')(z)| = \|\mathbb{B}V - \mathbb{B}V'\|_\infty \leq \gamma\epsilon$.

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.



(a) Patient treated as if "rationally" (b) Patient not treated "rationally" (c) Patient is diagnosed belatedly

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

Notation	Meaning	(first defined in)	Notation	Meaning	(first defined in)
ψ	problem setting	Section 3.1	s	environment state	Section 3.1
x	environment emission	Section 3.1	z	agent state, i.e. belief	Section 3.1
u	agent emission, i.e. action	Section 3.1	τ_{env}	environment transition	Section 3.1
τ	subjective transition	Section 3.1	ω_{env}	environment emission	Section 3.1
ω	subjective emission	Section 3.1	υ	utility (i.e. reward) function	Section 3.1
γ	discount factor	Section 3.1	ϕ	behavior	Section 3.1
ϕ_{demo}	demonstrated behavior	Section 3.2	ϕ_{imit}	imitation behavior	Section 3.2
θ	planning parameter	Section 3.1	θ_{norm}	normative parameter	Section 3.2
θ_{desc}	descriptive parameter	Section 3.2	π	decision policy	Section 3.1
ρ	recognition policy	Section 3.1	σ	specification policy	Section 4.1
F	forward planner	Section 3.1	G	inverse planner	Section 3.2
α^{-1}	flexibility coefficient	Section 4.2	β^{-1}	optimism coefficient	Section 4.2
η^{-1}	adaptivity coefficient	Section 4.2	$\tilde{\pi}$	action prior	Section 4.2
$\tilde{\sigma}$	model prior	Section 4.2	$\tilde{\varrho}$	belief prior	Section 4.2

References

[1] Aiping Li, Songchang Jin, Lumin Zhang, and Yan Jia. A sequential decision-theoretic model for medical diagnostic system. *Technology and Healthcare*, 2015.

[2] John A Clithero. Response times in economics: Looking through the lens of sequential sampling models. *Journal of Economic Psychology*, 2018.

[3] Jan Drugowitsch, Rubén Moreno-Bote, and Alexandre Pouget. Relation between belief and performance in perceptual decision making. *PloS one*, 2014.

[4] Gregory Wheeler. Bounded rationality. *SEP: Stanford Center for the Study of Language and Information*, 2018.

[5] Thomas L Griffiths, Falk Lieder, and Noah D Goodman. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, 2015.

[6] Tim Genewein, Felix Leibfried, Jordi Grau-Moya, and Daniel Alexander Braun. Bounded rationality, abstraction, and hierarchical decision-making: An information-theoretic optimality principle. *Frontiers in Robotics and AI*, 2015.

[7] Ned Augenblick and Matthew Rabin. Belief movement, uncertainty reduction, and rational updating. *UC Berkeley-Haas and Harvard University Mimeo*, 2018.

[8] Pedro A Ortega, Daniel A Braun, Justin Dyer, Kee-Eung Kim, and Naftali Tishby. Information-theoretic bounded rationality. *arXiv preprint*, 2015.

[9] L Robin Keller. The role of generalized utility theories in descriptive, prescriptive, and normative decision analysis. *Information and Decision Technologies*, 1989.

[10] Ludwig Johann Neumann, Oskar Morgenstern, et al. *Theory of games and economic behavior*. Princeton university press Princeton, 1947.

[11] Barbara A Mellers, Alan Schwartz, and Alan DJ Cooke. Judgment and decision making. *Annual review of psychology*, 1998.

[12] Yisong Yue and Hoang M Le. Imitation learning (presentation). *International Conference on Machine Learning (ICML)*, 2018.

[13] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. *International conference on Machine learning (ICML)*, 2004.

[14] Dean A Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural computation (NC)*, 1991.

[15] Michael Bain and Claude Sammut. A framework for behavioural cloning. *Machine Intelligence (MI)*, 1999.

[16] Umar Syed and Robert E Schapire. Imitation learning with a value-based prior. *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2007.

[17] Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. *International conference on artificial intelligence and statistics (AISTATS)*, 2010.

[18] Umar Syed and Robert E Schapire. A reduction from apprenticeship learning to classification. *Advances in neural information processing systems (NeurIPS)*, 2010.

[19] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. *International conference on artificial intelligence and statistics (AISTATS)*, 2011.

[20] Bilal Piot, Matthieu Geist, and Olivier Pietquin. Boosted and reward-regularized classification for apprenticeship learning. *International conference on Autonomous agents and multi-agent systems (AAMAS)*, 2014.

[21] Daniel Jarrett, Ioana Bica, and Mihaela van der Schaar. Strictly batch imitation learning by energy-based distribution matching. *Advances in neural information processing systems (NeurIPS)*, 2020.

[22] Alihan Hüyük, Daniel Jarrett, Cem Tekin, and Mihaela van der Schaar. Explaining by imitating: Understanding decisions by interpretable policy learning. *International Conference on Learning Representations (ICLR)*, 2021.

[23] Lionel Blondé and Alexandros Kalousis. Sample-efficient imitation learning via gans. *International conference on artificial intelligence and statistics (AISTATS)*, 2019.

[24] Ilya Kostrikov, Kumar Krishna Agrawal, Debidatta Dwibedi, Sergey Levine, and Jonathan Tompson. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation. *International Conference on Learning Representations (ICLR)*, 2019.

[25] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems (NeurIPS)*, 2016.

[26] Wonseok Jeon, Seokin Seo, and Kee-Eung Kim. A bayesian approach to generative adversarial imitation learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[27] Seyed Kamyar Seyed Ghasemipour, Richard Zemel, and Shixiang Gu. Understanding the relation of bc and irl through divergence minimization. *ICML Workshop on Deep Generative Models for Highly Structured Data*, 2019.

[28] Seyed Kamyar Seyed Ghasemipour, Richard Zemel, and Shixiang Gu. A divergence minimization perspective on imitation learning methods. *Conference on Robot Learning (CoRL)*, 2019.

[29] Liyiming Ke, Matt Barnes, Wen Sun, Gilwoo Lee, Sanjiban Choudhury, and Siddhartha Srinivasa. Imitation learning as f-divergence minimization. *arXiv preprint*, 2019.

[30] Liyiming Ke, Matt Barnes, Wen Sun, Gilwoo Lee, Sanjiban Choudhury, and Siddhartha Srinivasa. Imitation learning as f-divergence minimization. *International Workshop on the Algorithmic Foundations of Robotics (WAFR)*, 2020.

[31] Kee-Eung Kim and Hyun Soo Park. Imitation learning via kernel mean embedding. *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[32] Huang Xiao, Michael Herman, Joerg Wagner, Sebastian Ziesche, Jalal Etesami, and Thai Hong Linh. Wasserstein adversarial imitation learning. *arXiv preprint*, 2019.

[33] Robert Dadashi, Leonard Hussenot, Matthieu Geist, and Olivier Pietquin. Primal wasserstein imitation learning. *International Conference on Learning Representations (ICLR)*, 2021.

[34] Ilya Kostrikov, Ofir Nachum, and Jonathan Tompson. Imitation learning via off-policy distribution matching. *International Conference on Learning Representations (ICLR)*, 2020.

[35] Oleg Arenz and Gerhard Neumann. Non-adversarial imitation learning and its connections to adversarial methods. *arXiv preprint*, 2020.

[36] Srivatsan Srinivasan and Finale Doshi-Velez. Interpretable batch irl to extract clinician goals in icu hypotension management. *AMIA Summits on Translational Science Proceedings*, 2020.

[37] Xin Zhang, Yanhua Li, Ziming Zhang, and Zhi-Li Zhang. f-gail: Learning f-divergence for generative adversarial imitation learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[38] Nir Baram, Oron Anschel, and Shie Mannor. Model-based adversarial imitation learning. *arXiv preprint*, 2016.

[39] Nir Baram, Oron Anschel, and Shie Mannor. Model-based adversarial imitation learning. *International Conference on Machine Learning (ICML)*, 2017.

[40] Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. *International conference on Machine learning (ICML)*, 2000.

[41] Umar Syed and Robert E Schapire. A game-theoretic approach to apprenticeship learning. *Advances in neural information processing systems (NeurIPS)*, 2008.

[42] Umar Syed, Michael Bowling, and Robert E Schapire. Apprenticeship learning using linear programming. *International conference on Machine learning (ICML)*, 2008.

[43] Edouard Klein, Matthieu Geist, and Olivier Pietquin. Batch, off-policy and model-free apprenticeship learning. *European Workshop on Reinforcement Learning (EWRL)*, 2011.

[44] Takeshi Mori, Matthew Howard, and Sethu Vijayakumar. Model-free apprenticeship learning for transfer of human impedance behaviour. *IEEE-RAS International Conference on Humanoid Robots*, 2011.

[45] Donghun Lee, Srivatsan Srinivasan, and Finale Doshi-Velez. Truly batch apprenticeship learning with deep successor features. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

[46] Bilal Piot, Matthieu Geist, and Olivier Pietquin. Bridging the gap between imitation learning and irl. *IEEE transactions on neural networks and learning systems*, 2017.

[47] Edouard Klein, Matthieu Geist, Bilal Piot, and Olivier Pietquin. Irl through structured classification. *Advances in neural information processing systems (NeurIPS)*, 2012.

[48] Edouard Klein, Bilal Piot, Matthieu Geist, and Olivier Pietquin. A cascaded supervised learning approach to inverse reinforcement learning. *Joint European conference on machine learning and knowledge discovery in databases (ECML)*, 2013.

[49] Aristide CY Tossou and Christos Dimitrakakis. Probabilistic inverse reinforcement learning in unknown environments. *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2013.

[50] Vinamra Jain, Prashant Doshi, and Bikramjit Banerjee. Model-free irl using maximum likelihood estimation. *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[51] Gergely Neu and Csaba Szepesvári. Apprenticeship learning using irl and gradient methods. *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2007.

[52] Monica Babes, Vukosi Marivate, and Michael L Littman. Apprenticeship learning about multiple intentions. *International conference on Machine learning (ICML)*, 2011.

[53] Jonathan Ho, Jayesh Gupta, and Stefano Ermon. Model-free imitation learning with policy optimization. *International Conference on Machine Learning (ICML)*, 2016.

[54] Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. *International conference on machine learning (ICML)*, 2016.

[55] Matteo Pirotta and Marcello Restelli. Inverse reinforcement learning through policy gradient minimization. *AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

[56] Alberto Maria Metelli, Matteo Pirotta, and Marcello Restelli. Compatible reward inverse reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[57] Davide Tateo, Matteo Pirotta, Marcello Restelli, and Andrea Bonarini. Gradient-based minimization for multi-expert inverse reinforcement learning. *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017.

[58] Gergely Neu and Csaba Szepesvári. Training parsers by inverse reinforcement learning. *Machine learning (ML)*, 2009.

[59] Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

[60] Jaedeug Choi and Kee-Eung Kim. Map inference for bayesian irl. *Advances in Neural Information Processing Systems (NeurIPS)*, 2011.

[61] Christos Dimitrakakis and Constantin A Rothkopf. Bayesian multitask irl. *European workshop on reinforcement learning (EWRL)*, 2011.

[62] Constantin A Rothkopf and Christos Dimitrakakis. Preference elicitation and inverse reinforcement learning. *Joint European conference on machine learning and knowledge discovery in databases (ECML)*, 2011.

[63] Sreejith Balakrishnan, Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Harold Soh. Efficient exploration of reward functions in inverse reinforcement learning via bayesian optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[64] Ajay Kumar Tanwani and Aude Billard. Inverse reinforcement learning for compliant manipulation in letter handwriting. *National Center of Competence in Robotics (NCCR)*, 2013.

[65] McKane Andrus. Inverse reinforcement learning for dynamics. *Dissertation, University of California at Berkeley*, 2019.

[66] Stav Belogolovsky, Philip Korsunsky, Shie Mannor, Chen Tessler, and Tom Zahavy. Learning personalized treatments via irl. *arXiv preprint*, 2019.

[67] Sid Reddy, Anca Dragan, and Sergey Levine. Where do you think you're going?: Inferring beliefs about dynamics from behavior. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[68] Anirudha Majumdar, Sumeet Singh, Ajay Mandlekar, and Marco Pavone. Risk-sensitive inverse reinforcement learning via coherent risk models. *Robotics: Science and Systems*, 2017.

[69] Sumeet Singh, Jonathan Lacotte, Anirudha Majumdar, and Marco Pavone. Risk-sensitive inverse reinforcement learning via semi-and non-parametric methods. *International Journal of Robotics Research*, 2018.

[70] Jaedeug Choi and Kee-Eung Kim. Inverse reinforcement learning in partially observable environments. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2009.

[71] Jaedeug Choi and Kee-Eung Kim. Inverse reinforcement learning in partially observable environments. *Journal of Machine Learning Research (JMLR)*, 2011.

[72] Hamid R Chinaei and Brahim Chaib-Draa. An inverse reinforcement learning algorithm for partially observable domains with application on healthcare dialogue management. *International Conference on Machine Learning and Applications*, 2012.

[73] Ioana Bica, Daniel Jarrett, Alihan Hüyük, and Mihaela van der Schaar. Learning what-if explanations for sequential decision-making. *International Conference on Learning Representations (ICLR)*, 2021.

[74] Takaki Makino and Johane Takeuchi. Apprenticeship learning for model parameters of partially observable environments. *International Conference on Machine Learning (ICML)*, 2012.

[75] Daniel Jarrett and Mihaela van der Schaar. Inverse active sensing: Modeling and understanding timely decision-making. *International Conference on Machine Learning*, 2020.

[76] Kunal Pattanayak and Vikram Krishnamurthy. Inverse reinforcement learning for sequential hypothesis testing and search. *International Conference on Information Fusion (FUSION)*, 2020.

[77] Matthew Golub, Steven Chase, and Byron Yu. Learning an internal dynamics model from control demonstration. *International Conference on Machine Learning (ICML)*, 2013.

[78] Zhengwei Wu, Paul Schrater, and Xaq Pitkow. Inverse pomdp: Inferring what you think from what you do. *arXiv preprint*, 2018.

[79] Saurabh Daptardar, Paul Schrater, and Xaq Pitkow. Inverse rational control with partially observable continuous nonlinear dynamics. *arXiv preprint*, 2019.

[80] Minhae Kwon, Saurabh Daptardar, Paul Schrater, and Xaq Pitkow. Inverse rational control with partially observable continuous nonlinear dynamics. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[81] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. *AAAI Conference on Artificial Intelligence (AAAI)*, 2008.

[82] Abdeslam Boularias, Jens Kober, and Jan Peters. Relative entropy inverse reinforcement learning. *International conference on artificial intelligence and statistics (AISTATS)*, 2011.

[83] Mrinal Kalakrishnan, Peter Pastor, Ludovic Righetti, and Stefan Schaal. Learning objective functions for manipulation. *International Conference on Robotics and Automation (ICRA)*, 2013.

[84] Markus Wulfmeier, Peter Ondruska, and Ingmar Posner. Maximum entropy deep inverse reinforcement learning. *arXiv preprint*, 2015.

[85] Chelsea Finn, Paul Christiano, Pieter Abbeel, and Sergey Levine. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *NeurIPS Workshop on Adversarial Training*, 2016.

[86] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *International Conference on Learning Representations (ICLR)*, 2018.

[87] Ahmed H Qureshi, Byron Boots, and Michael C Yip. Adversarial imitation via variational inverse reinforcement learning. *International Conference on Learning Representations (ICLR)*, 2019.

[88] Paul Barde, Julien Roy, Wonseok Jeon, Joelle Pineau, Christopher Pal, and Derek Nowrouzezahrai. Adversarial soft advantage fitting: Imitation learning without policy optimization. *Advances in neural information processing systems (NeurIPS)*, 2020.

[89] Brian D Ziebart, J Andrew Bagnell, and Anind K Dey. Modeling interaction via the principle of maximum causal entropy. *International conference on Machine learning (ICML)*, 2010.

[90] Zhengyuan Zhou, Michael Bloem, and Nicholas Bambos. Infinite time horizon maximum causal entropy inverse reinforcement learning. *IEEE Transactions on Automatic Control (TACON)*, 2017.

[91] Kyungjae Lee, Sungjoon Choi, and Songhwai Oh. Maximum causal tsallis entropy imitation learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[92] Tien Mai, Kennard Chan, and Patrick Jaillet. Generalized maximum causal entropy for inverse reinforcement learning. *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[93] Michael Herman, Tobias Gindele, Jörg Wagner, Felix Schmitt, and Wolfram Burgard. Inverse reinforcement learning with simultaneous estimation of rewards and dynamics. *International conference on artificial intelligence and statistics (AISTATS)*, 2016.

[94] Michael Herman. Simultaneous estimation of rewards and dynamics in irl. *Dissertation, Albert-Ludwigs-Universitat Freiburg*, 2016.

[95] Layla El Asri, Bilal Piot, Matthieu Geist, Romain Laroche, and Olivier Pietquin. Score-based inverse reinforcement learning. *International conference on Autonomous agents and multi-agent systems (AAMAS)*, 2016.

[96] Benjamin Burchfiel, Carlo Tomasi, and Ronald Parr. Distance minimization for reward learning from scored trajectories. *AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

[97] Alexis Jacq, Matthieu Geist, Ana Paiva, and Olivier Pietquin. Learning from a learner. *International Conference on Machine Learning (ICML)*, 2019.

[98] Daniel S Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. *International Conference on Machine Learning (ICML)*, 2019.

[99] Daniel S Brown, Wonjoon Goo, and Scott Niekum. Better-than-demonstrator imitation learning via automatically-ranked demonstrations. *Conference on Robot Learning (CoRL)*, 2020.

[100] 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*, 2018.

[101] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. *International Conference on Machine Learning (ICML)*, 2017.

[102] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *International Conference on Machine Learning (ICML)*, 2018.

[103] Benjamin Eysenbach and Sergey Levine. If maxent rl is the answer, what is the question? *arXiv preprint*, 2019.

[104] Wenjie Shi, Shiji Song, and Cheng Wu. Soft policy gradient method for maximum entropy deep reinforcement learning. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

[105] Rohin Shah, Noah Gundotra, Pieter Abbeel, and Anca Dragan. Inferring reward functions from demonstrators with unknown biases. *OpenReview*, 2018.

[106] Rohin Shah, Noah Gundotra, Pieter Abbeel, and Anca D Dragan. On the feasibility of learning, rather than assuming, human biases for reward inference. *International Conference on Machine Learning (ICML)*, 2019.

[107] Jonathan Rubin, Ohad Shamir, and Naftali Tishby. Trading value and information in mdps. *Decision Making with Imperfect Decision Makers (Springer)*, 2012.

[108] Alexandre Galashov, Siddhant M Jayakumar, Leonard Hasenclever, Dhruva Tirumala, Jonathan Schwarz, Guillaume Desjardins, Wojciech M Czarnecki, Yee Whye Teh, Razvan Pascanu, and Nicolas Heess. Information asymmetry in kl-regularized rl. *International Conference on Learning Representations (ICLR)*, 2019.

[109] Mark K Ho, David Abel, Jonathan D Cohen, Michael L Littman, and Thomas L Griffiths. The efficiency of human cognition reflects planned information processing. *AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[110] Stas Tiomkin and Naftali Tishby. A unified bellman equation for causal information and value in markov decision processes. *arXiv preprint arXiv:1703.01585*, 2017.

[111] Felix Leibfried, Jordi Grau-Moya, and Haitham Bou-Ammar. An information-theoretic optimality principle for deep reinforcement learning. *NeurIPS Workshop on Deep Reinforcement Learning*, 2017.

[112] Matthew Hausknecht and Peter Stone. Deep recurrent q-learning for partially observable mdps. *AAAI Conference on Artificial Intelligence (AAAI)*, 2015.

[113] Pengfei Zhu, Xin Li, Pascal Poupart, and Guanghui Miao. On improving deep reinforcement learning for pomdps. *arXiv preprint*, 2017.

[114] Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep variational reinforcement learning for pomdps. *International Conference on Machine Learning (ICML)*, 2018.

[115] Amy Zhang, Zachary C Lipton, Luis Pineda, Kamyar Azizzadenesheli, Anima Anandkumar, Laurent Itti, Joelle Pineau, and Tommaso Furlanello. Learning causal state representations of partially observable environments. *arXiv preprint*, 2019.

[116] Dongqi Han, Kenji Doya, and Jun Tani. Variational recurrent models for solving partially observable control tasks. *arXiv preprint arXiv:1912.10703*, 2019.

[117] Joseph Futoma, Michael C Hughes, and Finale Doshi-Velez. Popcorn: Partially observed prediction constrained reinforcement learning. *International conference on artificial intelligence and statistics (AISTATS)*, 2020.

[118] Richard D Smallwood and Edward J Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations research*, 1973.

[119] Milos Hauskrecht. Value-function approximations for partially observable markov decision processes. *Journal of Artificial Intelligence Research (JAIR)*, 2000.

[120] Joelle Pineau, Geoff Gordon, Sebastian Thrun, et al. Point-based value iteration: An anytime algorithm for pomdps. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2003.

[121] Hanna Kurniawati, David Hsu, and Wee Sun Lee. Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces. *Robotics: Science and systems*, 2008.

[122] Mauricio Araya, Olivier Buffet, Vincent Thomas, and Françcois Charpillet. A pomdp extension with belief-dependent rewards. *Advances in Neural Information Processing Systems (NeurIPS)*, 2010.

[123] Mathieu Fehr, Olivier Buffet, Vincent Thomas, and Jilles Dibangoye. rho-pomdps have lipschitz-continuous epsilon-optimal value functions. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[124] F A Sonnenberg and J R Beck. Markov models in medical decision making: a practical guide. *Health Econ.*, 1983.

[125] C H Jackson, L D Sharples, S G Thompson, S W Duffy, and E Couto. Multistate Markov models for disease progression with classification error. *Statistician*, 2003.

[126] S E O'Bryant, S C Waring, C M Cullum, J Hall, L Lacritz, P J Massman, P J Lupo, J S Reisch, and R Doody. Staging dementia using Clinical Dementia Rating Scale Sum of Boxes scores: a Texas Alzheimer's research consortium study. *Arch. of Neurology*, 2008.

[127] D Jarrett, J Yoon, and M van der Schaar. Matchnet: Dynamic prediction in survival analysis using convolutional neural networks. *NeurIPS Workshop on Machine Learning for Health*, 2018.

[128] Daniel Jarrett, Jinsung Yoon, and Mihaela van der Schaar. Dynamic prediction in clinical survival analysis using temporal convolutional networks. *IEEE Journal of Biomedical and Health Informatics*, 2019.

[129] P Petousis, A Winter, W Speier, D R Aberle, W Hsu, and A A T Bui. Using sequential decision making to improve lung cancer screening performance. *IEEE Access*, 2019.

[130] F Cardoso, S Kyriakides, S Ohno, F Penault-Llorca, P Poortmans, I T Rubio, S Zackrisson, and E Senkus. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Anna. Oncology*, 2019.

[131] A M Alaa and M van der Schaar. Attentive state-space modeling of disease progression. *Advances in neural information processing systems (NeurIPS)*, 2019.

[132] X Wang, D Sontag, and F Wang. Unsupervised learning of disease progression models. *ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014.

[133] Clemens Heuberger. Inverse combinatorial optimization. *Journal of combinatorial optimization*, 2004.

[134] Kareem Amin and Satinder Singh. Towards resolving unidentifiability in inverse reinforcement learning. *arXiv preprint*, 2016.

[135] Kareem Amin, Nan Jiang, and Satinder Singh. Repeated inverse reinforcement learning. *Advances in neural information processing systems (NeurIPS)*, 2017.

[136] Stuart Armstrong and Sören Mindermann. Occam's razor is insufficient to infer the preferences of irrational agents. *Advances in neural information processing systems (NeurIPS)*, 2018.

[137] Paul Christiano. The easy goal inference problem is still hard. *AI Alignment*, 2015.

[138] Eric J Michaud, Adam Gleave, and Stuart Russell. Understanding learned reward functions. *NeurIPS Workshop on Deep Reinforcement Learning*, 2020.

[139] Adam Gleave, Michael Dennis, Shane Legg, Stuart Russell, and Jan Leike. Quantifying differences in reward functions. *International Conference on Learning Representations (ICLR)*, 2021.

[140] Daniel S Brown and Scott Niekum. Deep bayesian reward learning from preferences. *NeurIPS Workshop on Safety and Robustness in Decision-Making*, 2019.

[141] Daniel S Brown, Russell Coleman, Ravi Srinivasan, and Scott Niekum. Safe imitation learning via fast bayesian reward inference from preferences. *International Conference on Machine Learning (ICML)*, 2020.

[142] Nicolas Heess, David Silver, and Yee Whye Teh. Actor-critic reinforcement learning with energy-based policies. *European Workshop on Reinforcement Learning (EWRL)*, 2013.

[143] Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. *International Conference on Machine Learning (ICML)*, 2016.

[144] Nicolò Cesa-Bianchi, Claudio Gentile, Gábor Lugosi, and Gergely Neu. Boltzmann exploration done right. *Advances in neural information processing systems (NeurIPS)*, 2017.

[145] Amir Globerson, Eran Stark, Eilon Vaadia, and Naftali Tishby. The minimum information principle and its application to neural code analysis. *Proceedings of the National Academy of Sciences*, 2009.

[146] Naftali Tishby and Daniel Polani. Information theory of decisions and actions. *Perception-action cycle (Springer)*, 2011.

[147] Pedro A Ortega and Daniel A Braun. Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2013.

[148] Ian R Petersen, Matthew R James, and Paul Dupuis. Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *IEEE Transactions on Automatic Control*, 2000.

[149] Charalambos D Charalambous, Farzad Rezaei, and Andreas Kyprianou. Relations between information theory, robustness, and statistical mechanics of stochastic systems. *IEEE Conference on Decision and Control (CDC)*, 2004.

[150] Takayuki Osogami. Robustness and risk-sensitivity in markov decision processes. *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

[151] Jordi Grau-Moya, Felix Leibfried, Tim Genewein, and Daniel A Braun. Planning with information-processing constraints and model uncertainty in markov decision processes. *Joint European conference on machine learning and knowledge discovery in databases (ECML)*, 2016.

[152] Brian D Ziebart. Modeling purposeful adaptive behavior with the principle of maximum causal entropy. *Dissertation, Carnegie Mellon University*, 2010.

[153] Gerhard Kramer. Directed information for channels with feedback. *Dissertation, ETH Zurich*, 1998.

[154] James Massey. Causality, feedback and directed information. *International Symposium on Information Theory and Its Applications*, 1990.

[155] Hans Marko. The bidirectional communication theory-a generalization of information theory. *IEEE Transactions on Communications*, 1973.

[156] John B McKinlay, Carol L Link, et al. Sources of variation in physician adherence with clinical guidelines. *Journal of general internal medicine*, 2007.

[157] Matthias Bock, Gerhard Fritsch, and David L Hepner. Preoperative laboratory testing. *Anesthesiology clinics*, 2016.

[158] Jack W O'Sullivan, Carl Heneghan, Rafael Perera, Jason Oke, Jeffrey K Aronson, Brian Shine, and Ben Goldacre. Variation in diagnostic test requests and outcomes: a preliminary metric for openpathology. net. *Nature Scientific Reports*, 2018.

[159] Yunjie Song, Jonathan Skinner, Julie Bynum, Jason Sutherland, John E Wennberg, and Elliott S Fisher. Regional variations in diagnostic practices. *New England Journal of Medicine*, (1), 2010.

[160] Shannon K Martin and Adam S Cifu. Routine preoperative laboratory tests for elective surgery. *Journal of the American Medical Association (JAMA)*, 2017.

[161] M. Allen. Unnecessary tests and treatment explain why health care costs so much. *Scientific American*, 2017.

[162] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 1998.

[163] Razvan V Marinescu, Neil P Oxtoby, Alexandra L Young, Esther E Bron, Arthur W Toga, Michael W Weiner, Frederik Barkhof, Nick C Fox, Stefan Klein, Daniel C Alexander, et al. Tadpole challenge: Prediction of longitudinal evolution in alzheimer's disease. *arXiv preprint*, 2018.

[164] Edi Karni and Zvi Safra. Behavioral consistency in sequential decisions. *Progress in Decision, Utility and Risk Theory*, 1991.

[165] Kent Daniel, David Hirshleifer, and Avanidhar Subrahmanyam. Investor psychology and security market under-and overreactions. *The Journal of Finance*, 1998.

[166] Amos Tversky and Daniel Kahneman. Evidential impact of base rates. *Stanford University Department Of Psychology*, 1981.

[167] Charlotte L Allan and Klaus P Ebmeier. The influence of apoe4 on clinical progression of dementia: a meta-analysis. *International journal of geriatric psychiatry*, 2011.

[168] Sylvaine Artero, Marie-Laure Ancelin, Florence Portet, A Dupuy, Claudine Berr, Jean-François Dartigues, Christophe Tzourio, Olivier Rouaud, Michel Poncet, Florence Pasquier, et al. Risk profiles for mild cognitive impairment and progression to dementia are gender specific. *Journal of Neurology, Neurosurgery & Psychiatry*, 2008.

[169] Xue Hua, Derrek P Hibar, Suh Lee, Arthur W Toga, Clifford R Jack Jr, Michael W Weiner, Paul M Thompson, Alzheimer's Disease Neuroimaging Initiative, et al. Sex and age differences in atrophic rates: an adni study with n= 1368 mri scans. *Neurobiology of aging*, 2010.

[170] Brendan O'Donoghue, Remi Munos, Koray Kavukcuoglu, and Volodymyr Mnih. Combining policy gradient and q-learning. *International Conference on Learning Representations (ICLR)*, 2017.

[171] Momchil Tomov. Structure learning and uncertainty-guided exploration in the human brain. *Dissertation, Harvard University*, 2020.

[172] Tianwei Ni, Harshit Sikchi, Yufei Wang, Tejus Gupta, Lisa Lee, and Benjamin Eysenbach. F-irl: Inverse reinforcement learning via state marginal matching. *Conference on Robot Learning (CoRL)*, 2020.

[173] Hong Jun Jeon, Smitha Milli, and Anca D Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[174] Jeffrey Ely, Alexander Frankel, and Emir Kamenica. Suspense and surprise. *Journal of Political Economy*, 2015.

[175] Ahmed M Alaa and Mihaela van der Schaar. Balancing suspense and surprise: Timely decision making with endogenous information acquisition. *Advances in neural information processing systems (NeurIPS)*, 2016.

[176] Owain Evans and Noah D Goodman. Learning the preferences of bounded agents. *NeurIPS Workshop on Bounded Optimality*, 2015.

[177] Tan Zhi-Xuan, Jordyn L Mann, Tom Silver, Joshua B Tenenbaum, and Vikash K Mansinghka. Online bayesian goal inference for boundedly-rational planning agents. *Advances in neural information processing systems (NeurIPS)*, 2020.

[178] Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. Verifiable reinforcement learning via policy extraction. *Advances in neural information processing systems (NeurIPS)*, 2018.

[179] 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 (NeurIPS)*, 2020.

[180] Tom Bewley, Jonathan Lawry, and Arthur Richards. Modelling agent policies with interpretable imitation learning. *TAILOR Workshop at ECAI*, 2020.

[181] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[182] Tathagata Chakraborti, Sarath Sreedharan, Sachin Grover, and Subbarao Kambhampati. Plan explanations as model reconciliation: an empirical study. *International Conference on Human-Robot Interaction (HRI)*, 2019.

[183] Bradley Hayes and Julie A Shah. Improving robot controller transparency through autonomous policy explanation. *International Conference on Human-Robot Interaction (HRI)*, 2017.

[184] Sarath Sreedharan, Utkash Soni, Mudit Verma, Siddharth Srivastava, and Subbarao Kambhampati. Bridging the gap: Providing post-hoc symbolic explanations for sequential decision-making problems with black box simulators. *ICML Workshop on Human-in-the-Loop Learning*, 2020.

[185] Roy Fox and Naftali Tishby. Minimum-information lqg control part i: Memoryless controllers. *IEEE Conference on Decision and Control (CDC)*, 2016.

[186] Roy Fox and Naftali Tishby. Minimum-information lqg control part ii: Retentive controllers. *IEEE Conference on Decision and Control (CDC)*, 2016.

[187] Robert Babuska. Model-based imitation learning. *Springer Encyclopedia of the Sciences of Learning*, 2012.

[188] Yoshua Bengio and Paolo Frasconi. An input output hmm architecture. *Advances in neural information processing systems (NeurIPS)*, 1995.