Generalizing Off-Policy Evaluation From a Causal Perspective For Sequential Decision-Making

Sonali Parbhoo*, Shalmali Joshi* and Finale Doshi-Velez
Harvard University

September 2021

Abstract

Assessing the effects of a policy based on observational data from a different policy is a common problem across several high-stake decision-making domains, and several off-policy evaluation (OPE) techniques have been proposed for this purpose. However, these methods largely formulate OPE as a problem disassociated from the process used to generate the data (i.e. structural assumptions in the form of a causal graph). We argue that explicitly highlighting this association has important implications on our understanding of the fundamental limits of OPE. First, this implies that current formulation of OPE corresponds to a narrow set of tasks, i.e. a specific *causal estimand* which is focused on prospective evaluation of policies over populations or sub-populations. Second, we demonstrate how this association motivates natural desiderata to consider a more general set of *causal estimands*, particularly extending the role of OPE for *counterfactual* or *retrospective* off-policy evaluation at the level of individual units (e.g. patient-level) of the population. Further, a precise description of the causal estimand highlights which OPE estimands are identifiable from observational data under stated generative assumptions. For those OPE estimands that are not identifiable from observational data, the causal perspective further highlights where additional experimental data is *necessary* for identification, and thus naturally highlights situations where human expertise can aid identification and estimation. Furthermore, many formalisms of OPE overlook the role of *uncertainty* entirely in the estimation process.We demonstrate how specifically characterising the causal estimand highlights the different sources of uncertainty. The role of human expertise then naturally follows through in terms of managing the induced uncertainty. We discuss each of these aspects as actionable desiderata for future OPE research at scale and in-line with practical utility.

1 Introduction

In many real-world applications such as marketing [Silver et al., 2013], healthcare [Liao et al., 2019, Prasad et al., 2019, Parbhoo et al., 2017, 2018, Gottesman et al., 2019] and education [Mandel et al., 2014], it is common to reason about the effects of deploying one policy, based on data collected from a different policy. For instance, in healthcare, given the treatment history of a patient in ICU, we might be concerned about whether patient outcomes would improve if we changed the treatment protocol. Such problems have been posed as off-policy evaluation (OPE) tasks in reinforcement learning, and several methods have been proposed for performing OPE estimation.

Conventionally, OPE focuses on estimating the population and sub-population-level *prospective* performance of an evaluation policy based on data collected from a different behavior policy [Precup, 2000a, Sutton and Barto, 2018]. Evaluating prospective performance on (sub)-populations is akin to answering the question *"Will new patients have better outcomes if treated with A instead of B?"*. Yet in practice, we are also interested in questions of the form *"If we had acted differently, would patient X's condition have improved?"*. The two key distinctions are that these are inherently require reasoning about *retrospective* performance of an evaluation policy as well as require reasoning on an *individual* level [Heckman and Vytlacil, 2001, Richens et al., 2019, Valeri et al., 2016]. For prospective reasoning specifically at population and sub-population levels, there is a vast literature in identifying population and sub-population off-policy estimates, including characterizing their statistical properties [Dai et al., 2020, Precup, 2000b, Jiang and Li, 2016, Thomas and Brunskill, 2016, Levine et al., 2020]. Formalizing the latter (i.e. individual level, retrospective analysis) requires counterfactual reasoning. We argue that there is a need to broaden the traditional purview of OPE to include and enable individual-level reasoning, and formally, counterfactual analysis. We posit that i) this perspective is critical to understanding the limits of current ML-based modeling approaches for OPE, and ii) allows to generalize OPE to answer novel objectives of practical interest.

We argue that the unifying framework for OPE is to associate the OPE objective to the data-generating mechanism, more specifically, one endowed with an underlying Structural Causal Model (SCM). Using this framework, we can formalize OPE objectives as different *causal estimands* tied to appropriate assumptions over the structure and functional relationships in the associated SCM. Besides this unifying framework, there are several significant implications. First, whether and how well these causal quantities can be estimated from observational data (collected from a different policy) now reduces to figuring out whether the causal estimand is *observationally identifiable* (i.e., whether we can estimate it from observational data) given our structural assumptions. If identifiable, these necessary assumptions open the possibility of further understanding estimation properties (from a finite number of observational samples). This view also helps formalize OPE if an OPE estimate is non-identifiable from observational data. Specifically, non-identifiability can highlight the need for physical experiments or additional assumptions that could aid identification. Finally, formalizing the

*Equal contribution. Authors to prioritise ordering.

causal estimand can help determine the nature of domain expertise required to achieve multiple objectives of identification, estimation, and validating OPE estimates, which can critically improve the practical utility of OPE.

In this position paper, we discuss the need to formalize OPE through a causal lens to advance and generalize OPE to make it a useful tool for practice. Our objective is to use the causal perspective to chart a roadmap for generalizing OPE, in particular to expand and address individual level, sub-population level value estimates not just for prospective utility but also for retrospective reasoning in sequential decision-making settings. We first cover background focusing on a fundamental sequential decision-making setting of Markov Decision Process and Structural Causal Models in Section 2. We then show how the conventional view of OPE can be limiting and demonstrate how OPE tasks are formalised as causal estimands in Section 3. In Section 4, we discuss how outlining the appropriate causal estimand helps identify different sources of uncertainty, tied directly to i) identifiability of the OPE estimand from observational and/or experimental data, ii) make explicit the modeling assumptions and potential domain expertise required to estimate/ improve statistical properties of intermediate quantities of the OPE estimand. In doing so, we subsequently highlight that this characterization naturally formalizes the role of humans in generalizing OPE in Section 5. Specifically, we argue that human feedback may be necessary in terms of i) providing additional knowledge of the data-generating processes, including highlighting limits of where physical experimentation is possible, ii) explicitly conducting physical experiments to aid identifiability, iii) providing input that can improve statistical properties of OPE estimates in identifiable cases, or iv) providing domain knowledge about potential violations in the knowledge of the data-generating process, such as the amount of potential confounding, thereby enabling validation of OPE estimates. When assumptions about the underlying data-generating mechanisms are unclear or need to be relaxed, we argue that this organically generates a desiderata for developing methods that can help overcome various issues in OPE. We conclude by providing specific suggestions in the form of a roadmap for using OPE in practice in Section 6.

2 Background and Notation

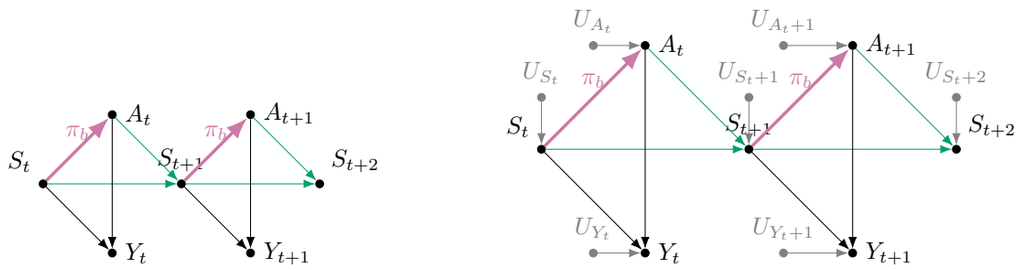
We begin by describing the notion of a Markov Decision Process (MDP) and its role in traditional off-policy evaluation. We subsequently show how augmenting the MDP with an SCM allows us to view different OPE objectives as *causal* estimands tied to data-generating mechanism of the MDP. We use the SCM formulation to formalize various estimands for OPE and discuss implications outlined above in subsequent sections. We use capital letters to denote random variables X and small letters to denote their realizations. Domain of a given random variable X is denoted by Ω_X. A collection of random variables is denoted by bold-faced caps letters $\mathbf{X}$.

Markov Decision-Process (MDP) [Bellman, 1957]. A canonical Data-Generating Process (DGP) assumed for OPE in reinforcement learning settings is the Markov Decision Process (MDP). MDP (finite or infinite horizon) is defined by the tuple $(\Omega_S, \Omega_A, \mathcal{P}, \mathcal{R}, p_0, \gamma)$ where Ω_S indicates the state-space, Ω_A indicates the action-space, P denotes the transition dynamics, i.e. $\mathcal{P} : \Omega_S \times \Omega_A \times \Omega_S \to [0, 1]$. The reward is denoted here by $Y \in \Omega_Y$. The reward function governing the MDP is denoted by $\mathcal{R} : \Omega_S \times \Omega_A \to \Omega_Y$, p_0 the initial state distribution and the discount factor γ. A policy provides a distribution over action given state i.e. functionally, $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$. Let $\mathcal{H}_T^\pi \triangleq \{S_0, A_1, Y_1, \cdots, S_{t-1}, A_t, Y_t, \cdots, S_{T-1}, A_T, Y_T\}$ be the trajectories collected under policy π and $\mathcal{H}_T^\pi$ is the collection of trajectories up to time T. Let $G(\mathcal{H}_T^\pi) = \sum_{t=0}^{T} \gamma^t Y_t$. A behavior policy is denoted with the subscript π_b and the evaluation policy is denoted by π_e. Figure 1(a) denotes the basic graphical model describing an MDP.

Next we cover background on Structural Causal Models (SCMs) and describe the SCM-augmented MDP that will subsequently be used to generalize OPE.

Structural Causal Models (SCM) [Pearl, 2009]. A structural causal model $\mathcal{M}$ describes the causal mechanisms driving a system. It consists of a tuple $\langle \mathbf{U}, \mathbf{V}, \mathbf{F}, \mathbf{P} \rangle$. Here $\mathbf{V}$ are the internal or endogenous variables in the causal system and $\mathbf{U}$ are the set of independent external or exogenous random variables that determine factors of variation in the system; The set of functions $\mathbf{F}$ govern the causal relationship between a variable $X \in \mathbf{V}$ and its causal parents Pa_X with the independent exogenous variation due to all exogenous variables $U_X \subseteq \mathbf{U}$ affecting X, given by $X \coloneqq f_X(Pa_X, U_X)$. The causal dependencies can be summarized in a Directed Acyclic Graph (DAG) $\mathcal{G}$ with $\mathbf{V}$ as the nodes and directed edges determined by the SCM. Exogenous nodes are unobserved. Any unobserved nodes between X and Z such that $U_X \cap U_Z \neq \varnothing$, induce dependence or confounding and are usually represented by a bidirectional edge between X and Z. The framework attributes probabilistic (Semi)-Markov assumptions to the joint distribution $\mathbf{P}$ over the nodes in the graph. This characterizes a probability distribution implying that we can observe samples corresponding to endogenous variables $\mathbf{V}$ true to the underlying causal graph and mechanism.

The formulation of structural causal models using the functional relationships and the endowed (Semi)-Markov distribution, allows us to define an intervention in the causal system. An atomic/hard intervention on a node anchors its value to a specific realization (e.g. assigning a treatment to a specific patient), thus removing the causal dependence on the node's parents and any stochasticity due to exogeneous variables. This is denoted as $do(X = x)$ for the node X and replaces the functional assignment $f_X(PA_X, U_X)$ with $X := x$. The induced distribution is known as an interventional distribution, denoted by $p(\mathbf{V}|do(X = x))$. The effect of this is that for all realisations inconsistent with $X = x$, the probability is 0. This allows for the reasoning: what would be the effect on $\mathbf{V} \setminus X$ if X is x, or practically, what would be the patient outcome if we only gave them an oral medicine. Note that interventions only affect descendant nodes. In OPE, we are often interested in identifying not just the effect of such a hard intervention, but of an alternative stochastic/deterministic policy, such



(a) (b)

Figure 1: MDP and MDP augmented with an SCM. (a): MDP with States S_t, Actions A_t and Outcomes/rewards Y_t sampled using behavior policy π_b. (b): SCM augmented MDP. Gray nodes here are exogenous latents that control the stochasticity of the corresponding node. Pink edges correspond to the behavior policy distribution. Green edges determine the transition dynamics in both figures.

as giving pills at a slightly lower frequency. These are known as *soft* interventions and result in modifying the functional assignment f_X with an alternative governing mechanism g_X, i.e. $X := g_X(Pa_X, U_X)$. A soft intervention is denoted by $\sigma(g_X)$. In this paper, we focus on both soft and hard interventions for a complete exposition.

Definition 1. *Soft Interventional Distribution: An intervention $\sigma(g_X)$ on a node X in an SCM $\mathcal{M}$ consists of replacing the governing mechanism of X given by the function $f_X(Pa_X, U_X)$ with a different governing causal mechanism $g_X(Pa_X, U_X)$ where Pa_X are the parents of X in a new DAG $\mathcal{G}^g$.*

An example of a soft intervention g is to evaluate for example, "what is the effect of treating a patient with pills over shots depending on symptoms?". Here the mechanism that recommends shots f is replaced by one recommending pills stochastically as a function of patient symptoms. This evaluation is indeed prospective. However, certain questions cannot be answered from such reasoning alone. Specifically, causal queries such as "would the patient's outcome be different if instead they were only provided with medications rather than an invasive procedure?", are *counterfactual* in nature. Counterfactual reasoning can be done for both soft and hard interventions, but requires inferring a model of the exogenous variables $P(\mathbf{U}|\mathbf{V} = \mathbf{v})$, that informs us of the likely stochasticity under the observational data. This can then be followed by an intervention with g on a causal system with exogenous noise priors $p(\mathbf{U})$ replaced by $p(\mathbf{U}|\mathbf{V} = \mathbf{v})$ (and is usually referred to as the abduction-step). Counterfactuals can therefore be used for *retrospective* reasoning in OPE.

Definition 2. *Counterfactual Distribution: Let $\mathcal{M}_{\mathbf{v}}$ correspond to the SCM where the exogenous noise model $p(\mathbf{U})$ in $\mathcal{M}$ is replaced by $p(\mathbf{U}|\mathbf{V} = \mathbf{v})$. Intervening with g on the resulting SCM induces the joint counterfactual distribution $P_{\mathcal{M}_{\sigma(g)|\mathbf{v}}}$. The corresponding counterfactual random variable(s) are denoted by $Y_t^{\sigma(g)}(\mathbf{U})$ where $\mathbf{U}$ indicates sampled units that are the target of counterfactual analysis. When clear from context we drop the $\sigma(\cdot)$ notation and refer to the variable as $Y_t^g(\mathbf{U})$.*

We can augment the MDP (Figure 1(a)) with an SCM to describe the underlying causal mechanisms governing the sequential decision-making system. The SCM analogue of an MDP can thus be re-stated formally and is denoted in Figure 1(b). State S_0 is a node without parents, and p_0 determines the stochasticity of S_0. The corresponding exogenous latent U_{S0} is such that $p(U_{S0}) = p_0$, and $S_0 = U_{S0}$ is the nominal functional relationship in the corresponding SCM. The policy π_b governs the relationship between actions A_t and its causal parent S_t for all time-steps. That is, $A_t = f_{A_t}(S_t, U_{At})$is governed by $\pi(A_t|S_t)$ where the parameterization of f and dependence on U_{At} is determined by the distribution π_b. For example, if A_t is normally distributed with mean S_t, and variance $\sigma^2_{A_t}$ i.e. $\pi(A_t|S_t) \triangleq \mathcal{N}(S_t, \sigma^2_{A_t})$, then a potential functional relation in the corresponding SCM is: $A_t = f_{At}(S_t, U_{At}) = S_t + U_{At}$ where $U_{At} \sim \mathcal{N}(0, \sigma^2_{A_t})$. Similarly the transition dynamics $\mathcal{P}$ governs the functional relationship $S_{t+1} = f_{S_{t+1}}(S_t, A_t, U_{St+1})$ implicitly. Note that f characterizes the deterministic component and all stochasticity is governed by exogenous U_{St}. Similarly for the reward function, i.e. $\mathcal{R}$ governs the functional relationship $Y_t = f_{Y_t}(S_t, A_t, U_{Yt})$. This perspective leads to an updated interpretation of the canonical OPE as a specific *causal estimand* in an MDP. An example of a soft intervention corresponding to an OPE task is the need to evaluate how an evaluation policy would perform by changing the mechanism governing A_t i.e. f_{A_t}, determined by π_b to an alternative function corresponding to the evaluation policy π_e). We formalize this in the next section.

3 Formalizing OPE as a causal estimand using SCMs

OPE is usually defined in the RL literature as that of estimating the cumulative reward of an evaluation policy π_e given observational trajectories from a different behavior policy π_b. More formally OPE is defined in Definition 3.

Definition 3. *Off-Policy Evaluation (OPE) [Precup, 2000a]: Off-policy evaluation estimates the performance $\mathbb{E}_{\pi_e}[G(\mathcal{H}_T^{\pi_e})|\mathcal{H}_T^{\pi_b}]$ of π_e based on a data set of trajectories $\mathcal{H}_T^{\pi_b} \triangleq \{S_0, A_1, Y_1, \cdots, S_t, A_t, Y_t, \cdots, S_{T-1}, A_T, Y_T\}$ each independently generated by different policy(ies) π_b.*

Definition 3 therefore is focused on i) expected reward from deploying π_e, and ii) implicitly focused on *prospective* performance of the policy (i.e. on new samples sampled from the data-generating mechanism that deploys π_e. This is made explicit by augmenting the MDP with an SCM. In the SCM framework, the canonical OPE objective in Definition 3 corresponds to a *soft intervention* of the form in Figure 2 (shown here

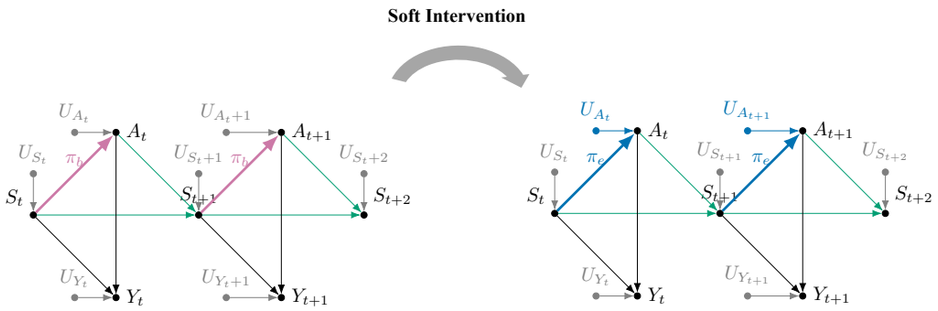


Figure 2: Example of an OPE task as a soft intervention in an MDP where the behavior policy π_b is replaced with evaluation policy π_e.

	Individual-Level	Subpopulation-Level	Population-Level
Counterfactual OPE (Retrospective)	E.g. Would patient i have survived had we changed the medication? $P(Y_T^{\pi_e}(U_i)\|\mathcal{H}_T^{\pi_b})$	E.g. Would female patients have survived had we changed the medication? $\mathbf{E}(Y_T^{\pi_e}(U_{female})\|\mathcal{H}_T^{\pi_b})$	E.g. Would all patients have survived had we changed the medication? $\mathbf{E}(Y_T^{\pi_e}(\boldsymbol{U})\|\mathcal{H}_T^{\pi_b})$
Interventional OPE (Prospective)	E.g. Will new patients improve if we changed the medication? $P(Y_T^{\pi_e}\|\mathcal{H}_T^{\pi_b})$	E.g. Will female patients improve if we changed the medication? $\mathbf{E}(Y_T^{\pi_e}\|\mathcal{H}_T^{\pi_b}, S_j = female)$	E.g. Will all patients survive if we change the medication? $\mathbf{E}(Y_T^{\pi_e}\|\mathcal{H}_T^{\pi_b})$

Canonical OPE task (Definition 1)

Figure 3: Formalizing OPE through a causal lens along two main axes: i) Estimand type and ii) Counterfactual vs Interventional OPE. Green shades indicate OPE tasks commonly considered in OPE reinforcement learning literature. Yellow shades indicate our proposed generalization by viewing OPE as a causal estimand. Each block contains an example task we might want to accomplish in OPE. Description of the appropriate causal estimand is provided in Section 3.

for the standard MDP case). This canonical definition corresponds to the causal estimand $\mathbb{E}[G(\mathcal{H}_T^{\pi_e})|\mathcal{H}_T^{\pi_b}]$ or when interested in the final outcome, just $\mathbb{E}[Y_T^{\pi_e}|\mathcal{H}_T^{\pi_b}]$.

Prospective OPE and interventional estimation. To make decisions about the utility of a policy at an individual level, we are not only interested in expected outcomes but on the probability distribution of the outcome of interest. For example, we may want to answer: *"Will new patients improve if we changed the medication?"*. The corresponding causal estimand is $p(Y_T^{\pi_e}|\mathcal{H}_t^{\pi_b})$ deviating from the canonical OPE estimand. Subgroup-level OPE estimation further demands conditional expectations based on specific attributes of interest: *"Will female patients improve if we changed the medication?"* and corresponds to the causal estimand $\mathbb{E}[Y_T^{\pi_e}|\mathcal{H}_T^{\pi_b}, S_{:,j} = female]$ where j indexes a specific attribute in the state representation vector (and is usually invariant over time hence not indexed by t. A more significant distinction which is excluded from the purview of OPE is whether one is interested in retrospective OPE i.e. answering whether for example *"Would this patient have survived had we changed the medication?"* This distinction is made clear in the SCM augmented MDP and the intuition of what it means to conduct such counterfactual reasoning.

First, any OPE task, requires a soft intervention for either retrospective or prospective analysis i.e. a using policy π_e instead of π_b, thus changing the mapping $S_t \to A_t$ (denoted in Figure 2). To sample actions, we this sample exogenous variables U_{At}'s according to the new policy mapping in either the retrospective or prospective OPE task. The main distinction is in the choice of units leveraged for inference over patient transitions and outcomes. Since the goal of interventional or prospective OPE tasks is to evaluate utility of a policy on *new* units or subpopulations post-soft-intervention, this inference is conducted on units re-sampled from the prior indicating prospective units. The OPE task can be individual-level, sub-population-level or population-level and can be estimated with appropriate conditioning post re-sampling. We make these distinctions clear in Figure 4. A unit corresponds to a sample associated with the MDP-SCM. The queries associated with this task are summarized in the bottom row of Figure 3. For individual level prospective OPE, we focus on the $P(Y_T^{\pi_e}|\mathcal{H}_T^{\pi_b})$, i.e. the prospective outcome after intervening with π_e using data collected from π_b. For sub-population-level estimands, we appropriately conditioned to identify units of interest (e.g. females) in Figure 3.

Retrospective OPE and counterfactual reasoning. On the other hand, for retrospective OPE, requiring counterfactual reasoning, the target unit(s) are fixed, identified by the task e.g. a fixed patient for individual-level reasoning, all females we have observed so far for sub-population counterfactual OPE and all units observed so far for population-level counterfactual OPE. To reason about the fixed patient then, the exogenous latents corresponding to the unit are held fixed (for unit i, we restrict to exogenous $\mathbf{U}_i$. Similarly for sub-population-level estimation (e.g. females, we restrict to female units denoted by $\mathbf{U}_{female}$ over which we take expectations post-intervention. The top row in Figure 3 highlight the corresponding causal estimand $(p(Y_T^{\pi_e}(U_i)|\mathcal{H}_T^{\pi_b})$ for individual-level counterfactual OPE and $\mathbb{E}_{\mathbf{U}_{female}}[Y_T^{\pi_e}(\mathbf{U}_{female})|\mathcal{H}_T^{\pi_b}]$ for sub-population-level counterfactual OPE). Figure 4 highlights how units are appropriately fixed, while the soft-intervention using π_e implies the actions will be sampled according to the intervened policy.

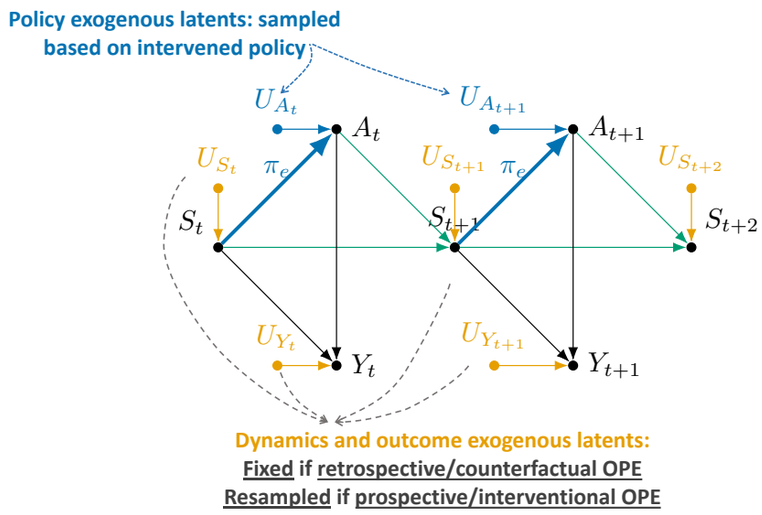


Figure 4: Distinguishing Interventional (prospective), and Counterfactual (retrospective) OPE, using distinct causal estimands. This figure depicts the distinction in the inference procedure (post-soft-intervention using the evaluation policy) for each case. For interventional OPE or prospective off-policy tasks, *new* units (samples from the MDP) are the target of inference. The appropriate estimand can be individual-level, sub-population-level or population-level (and can be appropriately conditioned on post re-sampling during inference). For counterfactual OPE or retrospective off-policy tasks, appropriate unit of inference (specific individual/unit) or a sub-group of individuals (sub-population units) are held *fixed* for inference.

In the following we focus on two tasks (one retrospective OPE and one prospective OPE) to demonstrate that effectively addressing all OPE tasks require assumptions on the data-generating process, outlining conditions that result in identifiability from observational data, which also further determine the intermediate quantities that may be required for estimation, and finally requirements for validation of OPE.

Example 1: Will all patients survive if we change the medication? This OPE task is corresponds to a population-level prospective OPE estimand (green box) in Figure 3. As described in previous sections, evaluating the (changed medication policy) based on data collected from (an old medication policy) corresponds to a soft intervention that modifies the policy mechanism. As a prospective task, inference requires sampling exogenous mechanisms from the prior that would allow to sample new units over which we take an expectation to get a population level estimate. Nonetheless, apriori, it is unclear whether observational data $\mathcal{H}_T^{\pi_b}$ is sufficient to estimate the corresponding causal estimand, $\mathbb{E}[\sum_{t=0}^{T} \gamma^t Y_t^{\pi_e} | \mathcal{H}_T^{\pi_b}]$ (generally, but also in the specific model shown in Figure 2).

We start with assuming that the data is generated from an MDP (Figure 2). We explain specific challenges using the example of Importance sampling (IS), which is often used to get an unbiased estimate of this estimand, when data is collected using policy π_b. Below we demonstrate the IS-OPE estimand [Precup, 2000b] and highlight assumptions made along the way that relate to and can be implied from assumptions on the data-generating mechanism.

$$\mathbb{E}_{\mathbf{S},\mathbf{Y}}[\sum_{t=0}^{T} \gamma^t Y_t^{\pi_e} | \mathcal{H}_T^{\pi_b}] = \mathbb{E}_{\mathbf{S},\mathbf{Y},A_t \sim \pi_e}[\sum_{t=0}^{T} \gamma^t Y_t^{A_t} | \mathcal{H}_T^{\pi_b}] \tag{1}$$

$$= \int \sum_{t=0}^{T} \gamma^t Y_t^{A_t} \prod_{t=1}^{T} p(Y_t|A_t,S_t)\pi_e(A_t|S_t)p(S_{t+1}|A_t,S_t)p(S_0) \tag{2}$$

$$= \int \sum_{t=0}^{T} \gamma^t Y_t \prod_{t=1}^{T} p(Y_t|A_t,S_t)\pi_e(A_t|S_t)p(S_{t+1}|A_t,S_t)p(S_0) \tag{3}$$

$$= \int \sum_{t=0}^{T} \gamma^t Y_t \prod_{t=1}^{T} p(Y_t|A_t,S_t)\pi_e(A_t|S_t)\frac{\pi_b(A_t|S_t)}{\pi_b(A_t|S_t)}p(S_{t+1}|A_t,S_t)p(S_0) \tag{4}$$

$$= \int \sum_{t=0}^{T} \gamma^t Y_t \prod_{t=1}^{T} p(Y_t|A_t,S_t)\pi_b(A_t|S_t)\frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)}p(S_{t+1}|A_t,S_t)p(S_0) \tag{5}$$

$$= \underbrace{\int \sum_{t=0}^{T} \gamma^t Y_t \prod_{t=1}^{T} p(Y_t|A_t,S_t)\pi_b(A_t|S_t)p(S_{t+1}|A_t,S_t)p(S_0)}_{\text{Trajectories under behavior policy}} \underbrace{\frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)}}_{\text{IS Re-weighting factor}} \tag{6}$$

Equation 6 shows importance re-weighting used on trajectories sampled from π_b in order to obtain OPE estimates corresponding to our target population-level estimand. From the derivation, we re-iterate and highlight assumptions Precup [2000a]. First, Equation 2 implies a factorization of the data-generating process which corresponds to the MDP assumption in Figure 2. Equation 3, replaces interventional random variables with observed counterparts due to sequential ignorability, which follows from the structural assumption of the

MDP and the backdoor-criterion [Richardson and Robins, 2013]: $Y_t^{a_t}, Y_t^{a_t'}, \cdots \perp A_t | \mathcal{H}_t$. Finally, Equation 6, is estimable due to assumptions of overlap: that the support of $\pi_e(A_t|S_t) > 0$ wherever $\pi_b(A_t|S_t) > 0$. More specifically, when we assume ignorability conditioning to S_t, that is, $Y_t^{a_t}, Y_t^{a_t'}, \cdots \perp A_t | S_t$, we assume that the underlying DGP corresponds to the MDP as in Figure 2 (left). When these assumptions hold, $\mathbb{E}[G(\mathcal{H}^{\pi_e})|do(\pi_e), \mathcal{H}_T^{\pi_b}]$ is estimable as we show above and IS provides an unbiased estimate of our OPE estimand.

Equation 6 therefore demonstrates how the causal estimand can be converted into a quantity estimable from observational data, albeit under the assumptions we highlighted. More importantly, we demonstrate that IS estimates of OPE are making implicit assumptions about observing all sources of confounding, overlap, and focusing on an expected measure of the value. When attempting prospective OPE evaluation for sub-groups, we further need to appropriately condition on (i.e. take expectations over sub-group identifying properties such as females). When interested in *individual-level* prospective OPE estimation, we may be interested in estimating not just the expected cumulative outcome but the complete distribution of the long-term come, thus changing the estimand. The issue of identifiability is implicit in the assumptions of sequential ignorability. When any parts of the state-space may be unobserved, this assumption may not be satisfied thus rendering IS-based prospective OPE estimates biased (see Figure 5 for an example and Section 5 on handling non-identifiability). We now turn a counterfactual OPE estimand to highlight how assumptions tied to the structure and functional relationships thereof can make identifiability (as well as estimation and validation) challenges transparent.

$$\mathbb{E}[Y_t^{A_t=a_t'}] = \sum_{a \in \{a_t, a_t'\}} \mathbb{E}[Y_t^{A_t=a_t'}|A_t = a]p(a) \tag{7}$$

Example 2: Would all patients have survived had we changed the medication? This OPE task, as pointed in out in Figure 3 (yellow box on top right), is a retrospective OPE task at the population-level. The most common example of this is the expected causal Effect of the Treatment on the Treated or ETT [Heckman, 1992], denoted by $\mathbb{E}[Y_t^{A_t=a_t'}|A_t = a_t]$ (where a_t is the original medication and a_t' is the changed medication. Specifically, this evaluates the expected outcome had the patients been treated with a_t' instead of a_t. Note that this is an example of a hard intervention. The estimand is explicitly conditioning on evidence that $A_t = a_t$ from observational data, but counterfactually allowing the treatment to vary (to a_t') to reason about the change in the mean. Another example, in the case of soft interventions, is when one may be interested in understanding the expected outcome using policy π_e having observed outcomes using policy π_b retrospectively, i.e. for the same set of patients treated with π_b denoted as: $\mathbb{E}[Y_T^{\pi_e}(\mathbf{U})|Y_T^{\pi_b}]$. We show the non-identifiability of ETT as an example here in the binary setting (note that this is a parametric assumption) as derived in [Shpitser and Pearl, 2012b] for the case where Y_t and A_t are binary in Equation 8. Note that Equation 7 can be written using law of total expectation. Then the target causal query is given by Equation 8.

$$\mathbb{E}[Y_t^{A_t=a_t'}|A_t = a_t] = \frac{\mathbb{E}[Y_t^{A_t=a_t'}] - \mathbb{E}[Y_t^{A_t=a_t'}|A_t = a_t']p(a_t')}{p(a_t)} \tag{8}$$

Thus to estimate ETT interventional quantities $\mathbb{E}[Y_t^{A_t=a_t'}]$ and $\mathbb{E}[Y_t^{A_t=a_t'}|a_t']$ are required, at least without further assumptions. Clearly, this estimand is not a function of observed variables directly, showing that such estimation is non-identifiable from historical data collected based on the previous medication. However if some experimental data is available, that allows us to anchor treatments to a_t', e.g. a randomized controlled trial, both of these quantities can be estimated, providing us with a way to counterfactually reason about the effect of alternative treatments on those who have been treated differently i.e. with a_t. In practice, some additional assumptions like monotonicity of the effects for binary treatment on the outcome can enable identifiability entirely from observational data and have been outlined in Pearl [2009, Chapter 9]. Identifiability with a general notion of monotonicity for discrete SCMs is outlined in Oberst and Sontag [2019]. These identifiability results again require ignorability/no unobserved confounding, thereby highlighting how these are closely tied to structural assumptions on the data-generating mechanisms. Further note that these identifiability results are applicable assuming treatments are binary and/or discrete. For other general estimands, it may be possible to get off-policy (i.e. from observational data) estimates of retrospective OPE estimands for specific data-generating assumptions, like no unobserved confounding and parametric assumptions like linearity.

These examples highlight that many (at times strong) assumptions are necessary to reliably solve OPE tasks, particularly retrospective OPE tasks. Challenges of identifiability can be further complex if the corresponding estimand is individual-level and require estimating a complete probability distribution as opposed to expectations. For the most general dynamic data-generating processes, identifiability of the appropriate counterfactual estimands without parametric assumptions i.e. entirely based on the graphical structure has been outlined in Shpitser and Pearl [2012a].

Formally, we are now in a position to define Counterfactual OPE and Interventional OPE. We focus on the most general OPE estimand (i.e. the probability distribution over target counterfactual/interventional quantity of interest) in each case and suggest that sub-population and/or population analogues can be correspondingly outlined. Note that purely in terms of the causal estimand, and the corresponding causal query, it is possible to unify these definitions (and notations) as is often the case in causality literature, we argue that there is significant merit in making these distinctions transparent, particularly for OPE tasks to explicitly contextualize contributions and ground the applicability of inferences made of OPE estimates.

Definition 4. *Counterfactual Off-Policy Evaluation: Counterfactual OPE corresponds to estimating the outcome distribution of a policy π_e based on a data set of trajectories $\mathcal{H}_T^{\pi_b} \triangleq \{S_0, A_1, Y_1, \cdots, S_{T-1}, A_T, Y_T\}$ generated independently by policy(ies) π_b on units* treated *with π_b corresponding to the causal estimand $P(G(\mathcal{H}_T^{\pi_e}(\mathbf{U}))|\mathcal{H}_T^{\pi_b})$, where $\mathbf{U}$ selects the appropriate target unit(s) of the counterfactual intervention π_e and $G(\cdot)$ is a function of counterfactual outcomes/trajectories.*

We now define Interventional OPE as follows:

Definition 5. *Interventional Off-Policy Evaluation: Interventional OPE corresponds to estimating the outcome distribution of a policy π_e based on a data set of trajectories $\mathcal{H}_T^{\pi_b} \triangleq \{S_0, A_1, Y_1, \cdots, S_{T-1}, A_T, Y_T\}$ generated independently by policy(ies) π_b on any prospective units treated with π_e corresponding to the causal estimand $P(G(\mathcal{H}_T^{\pi_e})|\mathcal{H}_T^{\pi_b})$ where $G(\cdot)$ is a function of interventional/prospective trajectories $\mathcal{H}_T^{\pi_e}$.*

More generally, the two examples and our definitions highlight gap in ML specific OPE literature suggesting OPE literature has largely focused on i) interventional analysis through off-policy evaluation, ii) focusing on estimates of the population, such as expectations, as opposed to individuals or rarely sub-populations and iii) assuming identifiability from observational data, thereby making tacit assumptions about the underlying data-generating procedure [Tennenholtz et al., 2019, Namkoong et al., 2020, Singh et al., 2021]. Applications like medical domains easily motivate the need for generalization along the lines summarized in Figure 3 as we are usually interested in evaluating patient level utility of policies, not just prospectively, but also retrospectively. In the following we discuss implications for OPE based on its formalization as a causal estimand. In particular, we propose that these distinctions be routinely highlighted, thereby clearly exposing the underlying assumptions and hence applicability of corresponding estimands [Tennenholtz et al., 2019, Singh et al., 2021, Namkoong et al., 2020, Oberst and Sontag, 2019]. We demonstrate the main utility of this exposition and the proposed generalizations specifically along these factors of OPE has practical implications. More specifically, these implications become clear by tying the associated estimation challenges to specific sources of uncertainties in the system.

4 Sources of Uncertainty in OPE.

Generalizing OPE as a causal estimand naturally highlights two key technical challenges for progress in OPE. These challenges are i) identifiability from available observational data, ii) estimation challenges. These challenges essentially manifest as induced uncertainty in the OPE estimation process. The causal characterization is useful because it further allows us to distinguish the nature of this uncertainty, and helps outline the role of humans in aiding and improving both the above aspects, thereby systematically mitigating the uncertainty in OPE estimation. Thus, these distinctions are not slight and have significant implications on our attempts to answer fundamental questions about OPE. We first discuss the different sources of uncertainty induced by the general characterization.

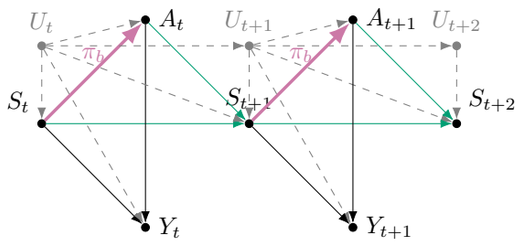


Figure 5: SCM analogue of a POMDP

1. Identifiability. Knowledge of the data-generating process, usually provided by human expertise, including functional relationships can determine whether OPE using data collected from a behavior policy π_b for another policy π_e is possible. Challenges like *partial observability* like in Figure 5 generally lead to non-identifiability for OPE, if we do not include additional assumptions. Non-identifiability implies the OPE estimate cannot be obtained even with infinite observational data from the behavior policy. This non-identifiability can be characterized as irreducible uncertainty in the OPE of interest. Characterizing this uncertainty received much less attention in ML literature for general causal inference Jesson et al. [2021a], let alone in OPE. Since Example 1 is a case of identifiable prospective OPE evaluation, there is no irreducible uncertainty induced due to non-identifiability. On the other hand, Example 2 suggests that without access to experimental data, the ETT estimate will have irreducible noise, and that experimentation should allow for estimating $\mathbb{E}[Y^{a_t}]$ to specifically target this source of uncertainty.

2. Estimation. Once the appropriate observational and experimental data required for identifiable OPE are established, we may either have sufficient observational data, or collect some additional experimental data to enable estimation. In practice, we often work with finite samples, which are in-turn used to estimate the appropriate intermediate quantities that provide the OPE estimate. For example, Equation 6, requires reasonable estimates of the dynamics model, and the behavior policy. In Example 2, we will obtain finite sample estimates of $\mathbb{E}[Y^{A_t=a_t}]$ from experimental data. Finite sample issues induce additional source(s) of uncertainty, in some cases requiring large amounts of observational/experimental data for accurate estimation. This is true irrespective of the type of estimand, and might put additional demands on data-collection needs. Since this uncertainty can be reduced by collecting additional data, this constitutes *modeling* or *epistemic* uncertainty.

Beyond structural assumptions about the data-generating process, parametric assumptions such as stationarity of MDP dynamics further add to the source of modeling or epistemic uncertainty. For instance, *non-stationary* exogenous variables often complicate model estimation, and can propagate uncertainty, thereby worsening how severe violations of assumptions like sufficient overlap [Namkoong et al., 2020, Joshi* et al., 2021] are in the estimation process. These assumptions are a form of parametric assumptions that leave some sources of uncertainty irreducible (by absorbing non-stationarity into irreducible noise). Finally, lack of sufficient data, or inappropriately accounting for confounding variables (i.e. state representations) can also introduce biases in the estimate depending on the severity of violations of assumptions of ignorability and overlap [Cinelli et al., 2020]. Issues of finite samples have only recently received attention and is necessary toward making causal inference a practical reality in ML and statistics [Jung et al., 2021, Chernozhukov et al., 2018]

Given these insights, we discuss dominant sources of uncertainty for different OPE estimands through our examples.

1. Example 1: This is an identifiable (from observational data assumed to be already available) interventional/prospective OPE task. Thus, the key sources of uncertainty here are modeling uncertainty associated with estimating intermediate quantities like dynamics estimation, and modeling the outcome $p(Y_t|S_t, A_t)$ from finite samples.

2. Example 2: This example of ETT requires both observational data to estimate the marginal $p(A_t)$, and experimental data to estimate $\mathbb{E}[Y_t^{A_t=a_t'}]$ each inducing different sources of modeling uncertainty depending on the amount and quality of available data. If only observational data is available, then this OPE estimand is not identifiable, and the irreducible or aleatoric uncertainty corresponds to that induced due to our uncertainty of the knowledge of $\mathbb{E}[Y_t^{A_t=a_t'}]$.

We now discuss the implication of characterizing these sources of uncertainty. First is the need to obtain additional domain knowledge to reduce uncertainty associated with identifiability. In traditional causality literature, the general focus of establishing counterfactual identifiability has been on graphical criteria with less emphasis on parametric assumptions on the corresponding functional relationships between covariates of interest. Of course there are several examples where making parametric assumptions can aid identifiability, even if non-parametric identifiability does not hold Hernán and Robins [2020], Angrist et al. [1996], Wright [1921]. Second is to resort to collecting experimental data, which in and of itself defeats the primary goal of OPE. More specifically, we discuss how humans can play a critical role for OPE where the causal estimand formalizes the type of human expertise and experimentation is required for practical OPE. We elaborate on these implications in detail in the following section.

5 The Role of Humans in OPE

In this section, we illustrate how the causal estimand, and explicitly outlining the sources of uncertainty helps to precisely characterize how human feedback should be incorporated and what the role can fill in improving the reliability of OPE. In particular, the role of incorporating human expertise can be useful for: i) providing modeling assumptions (and relaxations) under which we can describe the right causal OPE estimand, ii) physical experimentation to aid identifiability, iii) mitigating sources of uncertainty to compensate for limited data, and iv) providing assumptions that enable to understand the conditions for the validity of the OPE estimand. In essence, most of these principled ways of incorporating human feedback can be chalked up to reducing different sources of uncertainty and improving the reliability of OPE estimates. We now elaborate on these distinct roles that human feedback can fulfill in OPE, while explicitly discussing the kind of benefit it may provide.

Parametric assumptions and modeling constraints. In many situations, making additional modeling assumptions can be sufficient and valid to establish some form of identifiability, depending on the application. Human experts can assess the applicability of such assumptions to the domain, as well as provide any additional constraints. For example, assumptions such as consistency, monotonicity [Pearl, 2009, Chapter 9], or generalizations thereof are often used in epidemiological literature to estimate counterfactuals from observational data and have recently been explored in the context of OPE for model debugging [Oberst and Sontag, 2019]. In **Example 2**, assuming treatments are binary is a form of simple parametric assumption that can reduce the aleatoric uncertainty by reducing the number of interventional experiments that need to be conducted. These assumptions aid identifiability by making fundamental assumptions about the nature of counterfactuals, reducing aleatoric uncertainty. Additional knowledge like linearity and other functional assumptions, like the availability of noisy proxies of confounders [Miao et al., 2018, Tennenholtz et al., 2019], or availability of instrumental variables can help constrain the amount of online experimentation required for OPE [Xu et al., 2020]. These assumptions can help reduce modeling uncertainty, during estimation as well as reduce aleatoric uncertainty by constraining counterfactual distributions to specific functional forms.

When populations are heterogeneous, human input can additionally provide such functional assumptions at a higher level of granularity by drawing from existing clinical knowledge. If the estimated dynamics in models are unreliable or may shift across domains, domain knowledge about the amount of shift can improve the robustness of OPE estimates in practice [Singh et al., 2021]. In many cases, observational data may be insufficient, but experimentation is prohibitively costly or unethical. In such cases, domain experts can provide auxiliary model information like disease kinematics, or biophysical models collected from prior scientific experiments, as an alternative to simulating appropriate interventions [Adams et al., 2004, Ribba et al., 2012]. This form of human or expert feedback can therefore reduce modeling uncertainty in the system.

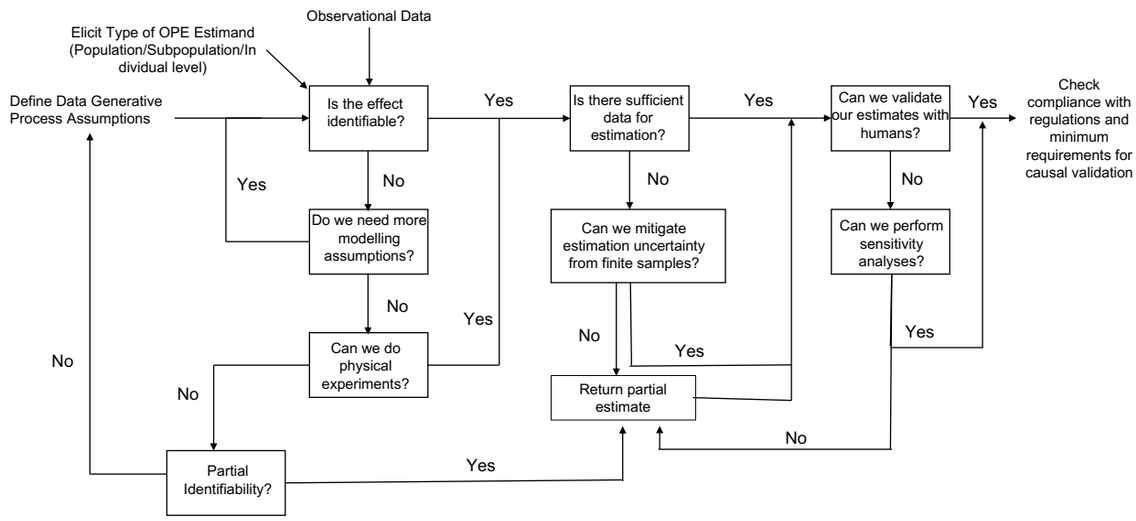


Figure 6: Roadmap for performing OPE in practice.

Most of these assumptions have been implicitly made to aid some form of identifiability in the most general sense and focused on population-level identifiability.

Physical experimentation for aiding identifiability. Establishing non-identifiability of a causal estimand corresponding to an OPE task can help identify the physical experimentation necessary to enable OPE for all cases we outline in Figure 3. As we saw in **Example 2**, OPE tasks requiring counterfactual reasoning, like ETT, might reveal the need for additional interventional/experimental data (or at the very least require domain expertise that further help establish identifiability). We demonstrated in Equation (8), that in this case, a combination of i) binary treatment and outcome (parametric assumption a human can provide and justify) as well as ii) interventional/experimental data is necessary when no additional assumptions like modeling assumptions can be justified. Without experimentation, there is irreducible uncertainty in the system. Further, consider the POMDP in Figure 5 where the prospective/retrospective OPE is unidentifiable (in expectation or for individual-level OPE). To aid identifiability, one primarily needs to collect unobserved confounders. Thus, physical experimentation need not directly imply expensive experiments like a randomized controlled trial, but could also help if additional confounding variables could be collected to establish identifiability. Data collection for specific variables can be less prohibitive than an active collection of interventional quantities and has been explored recently for (conditional) average treatment effect estimation [Wang et al., 2020, Jesson et al., 2021b]. Nonidentifiability exacerbated due to unobserved confounding and lack of the ability to experiment is the primary source of irreducible uncertainty or aleatoric uncertainty. Thus this form of human feedback can explicitly reduce aleatoric uncertainty. We argue that this is the main challenge where human input can further aid individual-level and retrospective OPE applications.

Mitigating estimation uncertainty from finite data. In many cases, while observational data collected from behavior policy might be sufficient for identifiability, the amount of data might be insufficient, significantly violating overlap assumptions which manifest in the increased variance of the OPE estimate. For instance, in Example 1, we may have very few samples collected from behavior policy, or the state representation may be high-dimensional. In these cases, humans may be able to analyze the validity of off-policy evaluation estimates *posthoc* to, for instance, reduce our modeling uncertainty. To this end, Gottesman et al. [2020] propose using humans to assess the validity of OPE estimates by manually analyzing the influential observations whose removal increases the estimation error. Similar approaches consider using human expertise to learn shaped control variates for variance reduction in OPE [Parbhoo et al., 2020] or using human expertise to define admissible rewards for high confidence OPE [Prasad et al., 2019]. This is another form of reducing stochasticity due to insufficient data i.e. reducing modeling uncertainty. While most existing work in this domain is applicable for prospective OPE tasks (and specifically **Example 1**), focused on sub-population/population-level OPE, there is significant potential for human input to reduce modeling uncertainty for retrospective and individual-level OPE tasks.

Validating OPE estimates. Validation of OPE is a significant but critical challenge for practical utility. The potential for human feedback is the least well explored especially for retrospective OPE. Technically, sensitivity analysis models are the main modeling frameworks that can allow the incorporation of domain knowledge that characterizes the severity of unobserved confounding in the system. For instance, it can help validate prospective as well as retrospective OPE tasks for complex settings such as the POMDP (Figure 5). Currently, the most commonly used sensitivity model known as the marginal sensitivity model, which constrains deviations to the propensity of treatment in the presence of unobserved confounders, and has been widely explored for population and sub-population-level prospective OPE in reinforcement learning [Kallus and Zhou, 2020, Namkoong et al., 2020]. Model debugging of this kind has been explored for counterfactual OPE for discrete SCMs for managing sepsis among diabetic patients in [Oberst and Sontag, 2019].

6 A Roadmap for OPE in Practice

In this work, we motivate the need to generalize the purview of OPE in sequential decision-making, specifically the need to distinguish prospective versus retrospective OPE tasks conducted at the individual, sub-population, and population level. To operationalize the generalization, we argue that OPE tasks should be framed in the context of the data generating process and a corresponding causal estimand. We discuss how this operationalization can highlight and help mitigate fundamental OPE challenges related to identification, estimation, and principally outline the role of human input to make OPE tasks practical. In doing so we highlight many open questions. In this section, we conclude by providing a prescriptive roadmap for addressing OPE tasks in practice, and technical challenges to consider along the way. Our prescriptive roadmap is summarized in Figure 6.

Elicit goals of OPE using a causal estimand. By eliciting the appropriate causal estimand, we can transparently identify whether an OPE task is prospective/retrospective and required at the level of individuals, sub-populations, and population levels. More importantly, this allows us to consider whether existing observational data collected from a behavior policy and additional assumptions can be justified to estimate the corresponding estimand. The corresponding assumptions in terms of the data-generating process (causal structure) and other assumptions (such as parametric assumptions) help provide precise specifications of the claim and validity of the OPE estimand.

Outline and justify nature of human input. The primary goal of OPE is to enable the evaluation of new policies based on observational data from behavior policies without additional experimentation. However, we highlight situations human input may be necessary and justified to solve specific OPE tasks. Here the form of human input can range from the need to conduct complex physical experimentation such as randomized controlled trials or in the form of data collection. This form of human input is explicitly aimed at mitigating non-identifiability issues. In some cases, human input might suffice to establish partial OPE estimates (bounds on the OPE estimate as opposed to point estimates). Even if an estimand is identifiable, we have to consider whether sufficient observational data is available to estimate the corresponding off-policy estimand. Here human input can help mitigate modeling uncertainty associated with estimation challenges by providing additional domain knowledge and would help improve the statistical properties of the corresponding estimators.

Identify minimal validation requirements using causal desiderata. The nature and type of validation required of OPE estimands are clear by focusing on the assumptions, the appropriate causal estimand, and the nature of human input incorporated to address identifiability and to obtain desirable estimation properties. In particular, these desiderata further identify the nature of validation, in the form of sensitivity analysis, specifically of the type that invalidates assumptions justified for the domain, the amount and nature of unobserved confounding, or structural assumptions to clearly identify granular conditions under which the OPE estimand is meaningful. Finally, to reliably deploy policies that are evaluated in this manner, it is imperative to consider additional human factors that will allow for appropriate interpretation of recommendations inferred from the OPE estimates under the specifications outlined by our desiderata. Note that most generally, assumptions of the data-generating process, as we have assumed throughout the draft, is a form of human expertise that allows to incorporate the most fundamental set of conditions of the domain under which the OPE estimand can be formulated. Using observational data as a means to learn such structures reliably is a nascent area called causal discovery or causal structure learning [Glymour et al., 2019].

In conclusion, these considerations should move us towards a more rigorous way of performing OPE in practice and create a better understanding of those factors essential for making OPE reliable for practice.

References

Brian M Adams, Harvey T Banks, Hee-Dae Kwon, and Hien T Tran. Dynamic multidrug therapies for hiv: Optimal and sti control approaches. *Mathematical Biosciences & Engineering*, 1(2):223, 2004.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

Richard Bellman. A markovian decision process. *Journal of mathematics and mechanics*, 6(5):679–684, 1957.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Carlos Cinelli, Andrew Forney, and Judea Pearl. A crash course in good and bad controls. *Available at SSRN*, 3689437, 2020.

Bo Dai, Ofir Nachum, Yinlam Chow, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. Coindice: Off-policy confidence interval estimation. *arXiv preprint arXiv:2010.11652*, 2020.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Omer Gottesman, Yao Liu, Scott Sussex, Emma Brunskill, and Finale Doshi-Velez. Combining parametric and nonparametric models for off-policy evaluation. *arXiv preprint arXiv:1905.05787*, 2019.

Omer Gottesman, Joseph Futoma, Yao Liu, Sonali Parbhoo, Leo Celi, Emma Brunskill, and Finale Doshi-Velez. Interpretable off-policy evaluation in reinforcement learning by highlighting influential transitions. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3658–3667. PMLR, 13–18 Jul 2020. URL `http://proceedings.mlr.press/v119/gottesman20a.html`.

James J Heckman. Randomization and social policy evaluation. *Evaluating welfare and training programs*, 1: 201–30, 1992.

James J Heckman and Edward Vytlacil. Policy-relevant treatment effects. *American Economic Review*, 91(2): 107–111, 2001.

Miguel A Hernán and James M Robins. Causal inference: what if, 2020.

Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. *arXiv preprint arXiv:2103.04850*, 2021a.

Andrew Jesson, Panagiotis Tigas, Joost van Amersfoort, Andreas Kirsch, Uri Shalit, and Yarin Gal. Causalbald: Deep bayesian active learning of outcomes to infer treatment-effects from observational data, 2021b.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR, 2016.

Shalmali Joshi*, Sonali Parbhoo*, and Finale Doshi-Velez. Interpretable learning-to-defer for sequential decision-making. In *ICML Workshop for Interpretable Machine Learning for Healthcare*, 2021.

Yonghan Jung, Jin Tian, and Elias Bareinboim. Estimating identifiable causal effects through double machine learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021.

Nathan Kallus and Angela Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. *arXiv preprint arXiv:2002.04518*, 2020.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Peng Liao, Predrag Klasnja, and Susan Murphy. Off-policy estimation of long-term average outcomes with applications to mobile health. *arXiv preprint arXiv:1912.13088*, 2019.

Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1077–1084. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. *arXiv preprint arXiv:2003.05623*, 2020.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, pages 4881–4890. PMLR, 2019.

Sonali Parbhoo, Jasmina Bogojeska, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. Combining kernel and model based learning for hiv therapy selection. *AMIA Summits on Translational Science Proceedings*, 2017:239, 2017.

Sonali Parbhoo, Omer Gottesman, Andrew Slavin Ross, Matthieu Komorowski, Aldo Faisal, Isabella Bon, Volker Roth, and Finale Doshi-Velez. Improving counterfactual reasoning with kernelised dynamic mixing models. *PloS one*, 13(11):e0205839, 2018.

Sonali Parbhoo, Omer Gottesman, and Finale Doshi-Velez. Shaping control variates for off-policy evaluation. In *Offline Reinforcement Learning Workshop at Neural Information Processing Systems (NeurIPS)*, 2020.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Niranjani Prasad, Barbara E Engelhardt, and Finale Doshi-Velez. Defining admissible rewards for high confidence policy evaluation. *arXiv preprint arXiv:1905.13167*, 2019.

Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000a.

Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000b.

Benjamin Ribba, Gentian Kaloshi, Mathieu Peyre, Damien Ricard, Vincent Calvez, Michel Tod, Branka Čajavec-Bernard, Ahmed Idbaih, Dimitri Psimaras, Linda Dainese, et al. A tumor growth inhibition model for low-grade glioma treated with chemotherapy or radiotherapy. *Clinical Cancer Research*, 2012.

Thomas S Richardson and James M Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. Counterfactual diagnosis. *arXiv preprint arXiv:1910.06772*, 2019.

Ilya Shpitser and Judea Pearl. What counterfactuals can be tested. *arXiv preprint arXiv:1206.5294*, 2012a.

Ilya Shpitser and Judea Pearl. Effects of treatment on the treated: Identification and generalization. *arXiv preprint arXiv:1205.2615*, 2012b.

David Silver, Leonard Newnham, David Barker, Suzanne Weller, and Jason McFall. Concurrent reinforcement learning from customer interactions. In *International Conference on Machine Learning*, pages 924–932, 2013.

Harvineet Singh, Shalmali Joshi, Finale Doshi-Velez, and Himabindu Lakkaraju. Learning under adversarial and interventional shifts. *arXiv preprint arXiv:2103.15933*, 2021.

R.S. Sutton and A.G. Barto. *Reinforcement Learning, second edition: An Introduction*. Adaptive Computation and Machine Learning series. MIT Press, 2018. ISBN 9780262352703. URL `https://books.google.com/books?id=uWVODwAAQBAJ`.

Guy Tennenholtz, Shie Mannor, and Uri Shalit. Off-policy evaluation in partially observable environments. *arXiv preprint arXiv:1909.03739*, 2019.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148. PMLR, 2016.

Linda Valeri, Jarvis T Chen, Xabier Garcia-Albeniz, Nancy Krieger, Tyler J VanderWeele, and Brent A Coull. The role of stage at diagnosis in colorectal cancer black–white survival disparities: a counterfactual causal inference approach. *Cancer Epidemiology and Prevention Biomarkers*, 25(1):83–89, 2016.

Shirly Wang, Seung Eun Yi, Shalmali Joshi, and Marzyeh Ghassemi. Confounding feature acquisition for causal effect estimation. In *Machine Learning for Health*, pages 379–396. PMLR, 2020.

Sewall Wright. Correlation and causation. *Journal of Agricultural Research*, 1921.

Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. Learning deep features in instrumental variable regression. *arXiv preprint arXiv:2010.07154*, 2020.