Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Adapting to drift you cannot see: passive identifiability decides whether online regime inference pays off in epidemic control

**Tian-Yu Wang**

wangty0709@hotmail.com

Guangdong Peizheng College https://orcid.org/0000-0002-5124-7965

**Research Article**

**Keywords:** meta-reinforcement learning; epidemic intervention; regime drift; identifiability; dual control; infectious disease modelling

**Posted Date:** November 11th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11679441/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** The authors declare no competing interests.

# Adapting to drift you cannot see: passive identifiability decides whether online regime inference pays off in epidemic control

**Short title:** Online regime inference for meta-RL in epidemics

**Author:** Tian-Yu Wang (ORCID: 0000-0002-5124-7965)

**Affiliation:** Medical Office, General Affairs Office, Guangdong Peizheng College, Guangzhou 510830, Guangdong, China

**Email:** wangty0709@hotmail.com

**Keywords:** meta-reinforcement learning; epidemic intervention; regime drift; identifiability; dual control; infectious disease modelling

## Abstract

Public-health decisions about when to intervene against a respiratory pathogen are made repeatedly under conditions that move. Meta-reinforcement learning is built for fast adaptation, and epidemics drift, so an adaptive policy looks like the right tool. Whether continuous re-inference of the latent regime actually improves decisions has not been tested against a design that can separate the two reasons it might not. We built a seasonal SEIRS-type environment with a reporting layer in which a seven-dimensional regime drifts while the agent observes only reported case counts, together with three families that distinguish whether a decision-relevant difference exists from whether it can be recovered without acting: fixed intervention parameters (passive), parameters affine in the passively visible transmission rate (linked), and parameters that act only through the intervention (action-revealed). A PEARL-style context encoder trained with PPO was compared with frozen inference, an oracle, and non-adaptive baselines across 25 seeds. Continuous re-inference gave no reliable advantage over an estimate frozen at the start of an episode in either family (+0.32 linked, +0.02 action-revealed, across-seed standard deviations 0.47 and 2.96). Knowing the true regime was worth +0.06 in the linked family against +2.03 in the action-revealed family, so the value of the regime concentrates where it cannot be read passively. A passive identifiability index, computed from intervention-free trajectories before any training, separates the families out of sample (R2 0.591 against 0.158) and serves as a design-time test: where the parameters that set the best action are already readable from passive surveillance, a fixed policy does as well. A gate signal built from the agent's own model carried information about harmful adaptation steps but did not beat a plain predictive-variance baseline.

## 1 Introduction

Respiratory pathogens present public-health systems with a decision problem that does not stop: whether to intervene, at what strength, and when to relax. The answer is not a property of the pathogen alone. Reviews of artificial intelligence for infectious-disease modelling have emphasised that the difficult part of this class of problems is not forecasting alone but choosing actions under conditions that change and are only partly observable [1]. Influenza and related viruses circulate with a pronounced seasonal cycle [2,3], so the same intervention can avert a large burden in one month and

buy little in another, and the effect of an intervention depends on how a population responds to it, which is itself uncertain.

Reinforcement learning (RL) has been applied to epidemic control with promising results. Policies have been trained to control influenza in metapopulation models [4], to select COVID-19 strategies [5], to derive control policies for outbreaks [6], to target mobility interventions under unobservable state [7], and to allocate scarce public-health resources with contextual bandits [8]. Most of this work trains a policy inside a single, fixed epidemiological environment. In practice the environment is not fixed. New variants, changes in contact behaviour, revisions to testing and reporting, and the accumulation of immunity all alter the mapping from what is observed to the action that is best. A policy tuned to one parameterisation must be retrained when the parameterisation moves. Transfer RL for partially observable, time-varying epidemic parameters is one response to this, and it makes the difficulty explicit: the parameters that determine the best action are not known and they change [9].

Meta-reinforcement learning (meta-RL) is built for this setting. Rather than solving one task, a meta-agent is trained over a distribution of tasks so that it can adapt to a new one from a small amount of experience [10,11,12]. Probabilistic approaches such as PEARL condition a policy on a latent context inferred from recent transitions [13], and Bayes-adaptive methods such as VariBAD infer a posterior over task parameters and act under it [14]. The area is surveyed in a recent tutorial [15]. Extensions handle tasks that are non-stationary within an episode [16] and learn task representations without supervision [17]. Classical control theory recognised the same tension decades earlier: an action serves both to control the system and to reveal its parameters, and the two objectives can conflict [18,19]. Epidemic intervention is an unusually sharp case of that conflict, because exploration is not free of consequence: an action taken to learn how a population responds is an action taken against that population. Despite this, meta-RL has rarely been applied to epidemic intervention, and it is not obvious when it should help.

A natural expectation is that explicit online regime inference will help whenever the latent situation drifts: if the environment moves, re-inferring it should pay. That expectation leaves out a precondition. What decides the value of online inference is not how much the environment drifts but whether the parameters that matter for the decision can be recovered from passive observation. If those parameters leave a signature on the case trajectory even when nothing is done, an agent can read them from that trajectory, and re-inference adds little. If they act on the dynamics only through the intervention itself, no amount of intervention-free observation separates two tasks that differ in them. An agent that does not act on the ambiguity cannot resolve it. An agent that does act pays, in the same step, for an action that may be wrong.

We study this distinction in a controlled simulation and make three contributions. The first is a construction. We build three environment families that separate the existence of a decision-relevant difference from its identifiability. In the passive family the intervention parameters are fixed, so there is little to gain. In the linked family they vary but are affine functions of the passively visible transmission rate, so a difference exists and is readable from passive case trajectories. In the action-revealed family they vary and enter the dynamics only through the action term, so that two such tasks generate identical

observations until the agent intervenes. Comparing the linked and action-revealed families isolates identifiability from the mere existence of a difference, with the passive family as a reference in which there is nothing to gain. The construction yields two falsifiable predictions, both stated on the post-drift window in which adaptation is visible: that the difference in post-drift captured fraction between the online and frozen agents is close to zero in the linked family and positive in the action-revealed family; and that the additional value of knowing the true regime, the oracle-minus-online difference, is much smaller in the linked family than in the action-revealed family.

The second contribution turns the requirement to act into quantities that can be measured before anything is trained: the passive identifiability index (PII), the out-of-sample accuracy with which the best constant intervention can be predicted from passive trajectories alone, and a gate signal computed from the agent's own model, requiring no regime labels, that is intended to flag the steps at which adaptation is harmful.

The third is the outcome of running that test. We report boundary conditions, not a ranking of methods: the extra value of the true regime was +2.03 in the action-revealed family against +0.06 in the linked family, continuous re-inference gave no reliable advantage over a frozen estimate in either family, and the PII separated the two families before any meta-training. The remainder of the paper describes the environment and agents (Materials and Methods), the boundary condition and the evaluation of the index and the gate signal (Results), and the interpretation and limits of the findings (Discussion).

## 2 Results

### 2.1 The simulated drift is of the same magnitude as drift in real respiratory surveillance

We first asked whether the drift the agents face is of a magnitude that occurs in real surveillance, so that the comparison is not decided by an artificial mechanism. We took real respiratory-pathogen series [20,21,22] and applied the same change-point rule to them and to the simulator. The calibration is the subject of Fig 1; the exact numbers are given in electronic supplementary material, S6 Appendix.

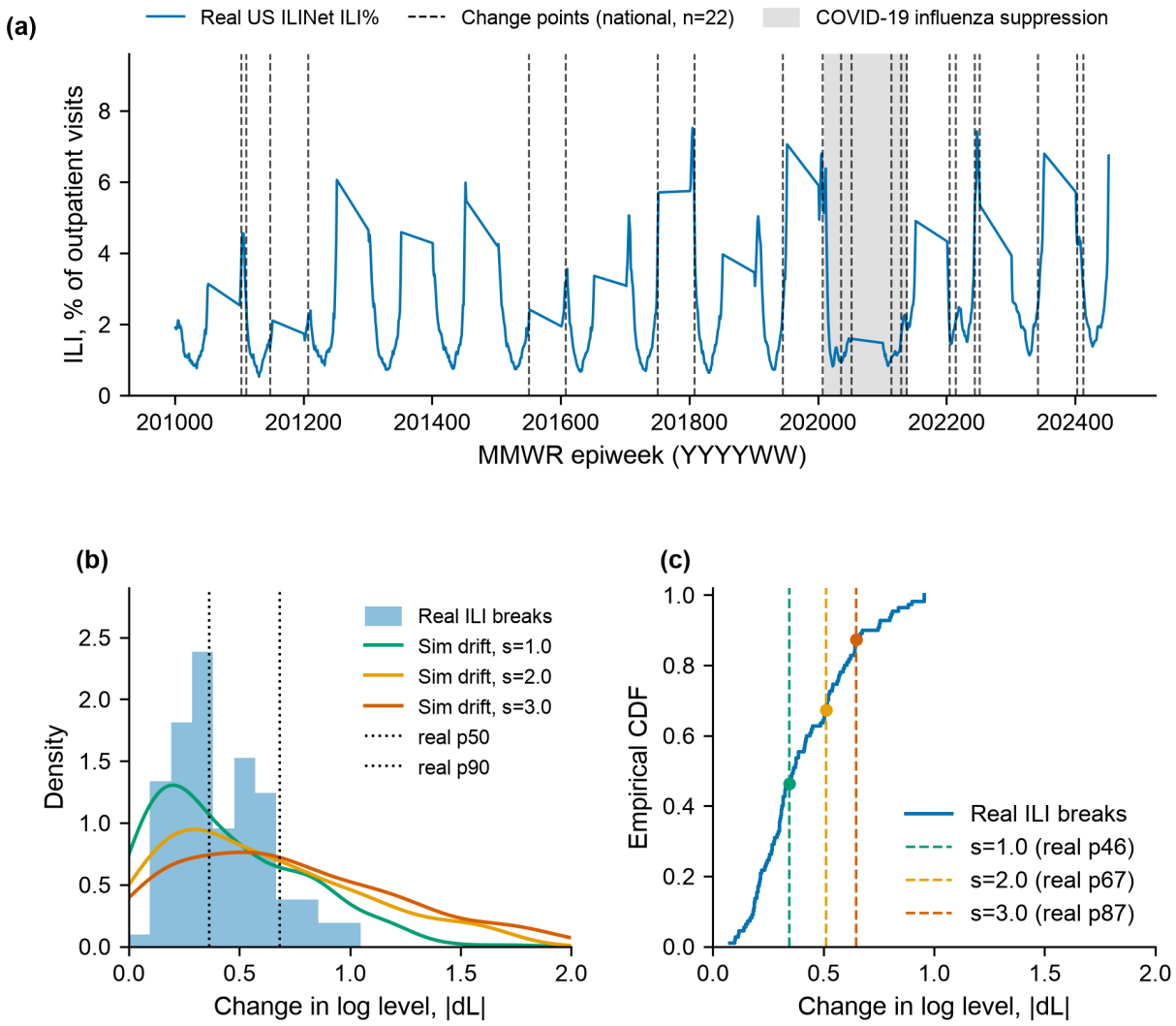



*Fig. 1.* Real respiratory-surveillance drift matches the simulated drift. (a) Weekly US ILINet ILI% with detected level change points (dashed) and the COVID-19 suppression interval (shaded). (b) Density of real ILI level changes versus simulated drift at strengths 1.0, 2.0 and 3.0. (c) Empirical CDF of real ILI level changes; markers are the simulated medians, which fall at real percentiles 46.4%, 67.3% and 87.3%. Data sources: Delphi Epidata FluView (US ILINet) and the additional data streams listed in Materials and Methods.

**2.2 Online regime inference pays only when the regime is not passively identifiable**

We then compared the agents across the three environment families, in the post-drift window in which adaptation is visible. The predicted pattern is a boundary condition, not a uniform gain: continuous re-inference should be worth little where the regime is passively readable and a measurable amount where it is not. The data support a boundary condition, but they locate it in the oracle contrast rather than in the frozen contrast.

Against a context estimate frozen at the start of the episode, the online agent's post-drift captured fraction changed by +0.32 in the linked family and by +0.02 in the action-revealed family (across-seed standard deviations 0.47 and 2.96). Continuous re-inference therefore gave no reliable advantage over a frozen estimate in either family. Against an oracle given the true regime, the extra value of that

knowledge was +0.06 in the linked family against +2.03 in the action-revealed family. That contrast is heavy-tailed across seeds, because a small number of cells draw a near-degenerate available gap and the captured ratio then divides by a near-zero denominator (Section 3.1). Its median is -0.06 in the linked family, where 9 of 25 seeds are positive, and +0.22 in the action-revealed family, where 17 of 25 are (one-sided sign test, p = 0.054). The action-revealed value exceeds the linked value in 18 of 25 seeds (p = 0.022), with the same ordering when the contrast is computed in raw post-drift return (17 of 25 seeds).

What separates the families is the value of the regime's information, not the advantage of re-inferring it continuously. Where the regime is readable, the online agent already recovers essentially all of what the oracle knows. Where it is not, neither agent recovers much of it. The online agent also exceeded the non-adapting baseline in the two families where the captured fraction is defined (Fig 2), but that comparison is not the informative one here, because the domain-randomised baseline is weak in this environment (Section 3.1).

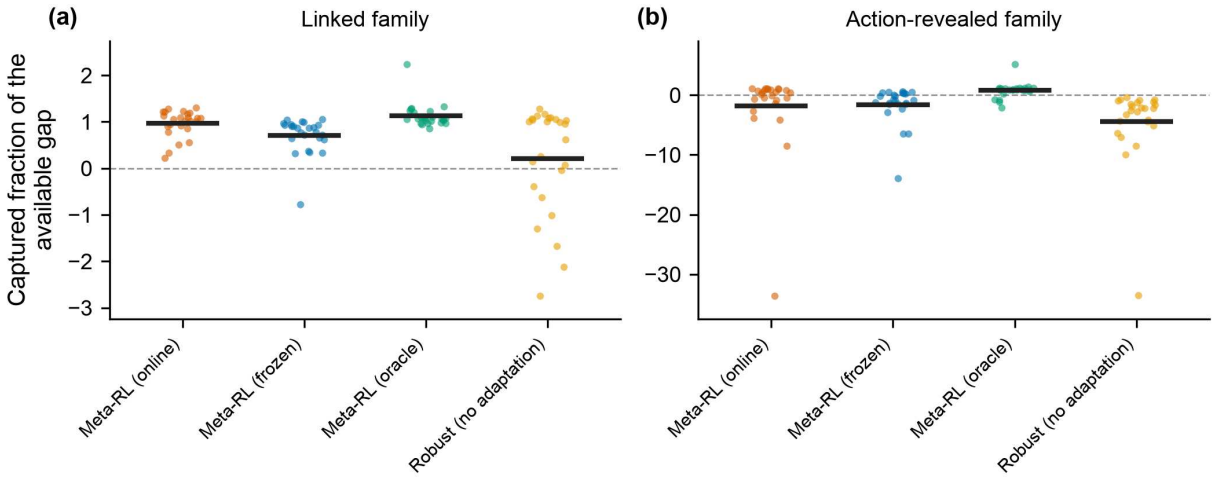



*Fig. 2.* Online inference pays off only where the regime cannot be read passively. Two panels: (a) linked family, (b) action-revealed family. Both plot the captured fraction over the full episode (share of the per-task available gap; definition in Materials and Methods), averaged over the five non-stationary drift kinds at strength 1.0. Each point is one training seed and the horizontal bar is the group mean; the groups are, left to right, the online agent (0.97, -1.83), the frozen ablation (0.71, -1.59), the oracle (1.13, 0.81) and the non-adapting baseline (0.20, -4.38), where each series has one value per panel. The panels use independent vertical scales, because a small number of seeds in panel (b) draw a near-degenerate available gap (Section 3.1). The oracle ranks first and the non-adapting baseline last in both panels, and the same holds in the post-drift window, which is the window in which the contrasts quoted in the text are computed (S6 Appendix); the online and frozen agents differ by at most 0.03 captured units in the post-drift window of panel (b). The passive family is not plotted: its available gap is only 2.53, so the captured-fraction ratio is numerically unstable there (Section 3.1), and its raw returns are tabulated in S3 Table. The available gaps shown are the experimental-protocol values (S6 Appendix): (a) 31.12, (b) 25.33.

**2.3 A model-derived gate signal carries information but does not beat a simpler baseline**

A defender of adaptation needs to know when to trust it. We evaluated whether a signal computed online from the agent's own context model, without regime labels, can separate the steps at which adaptation is harmful from those at which it helps, and we compared it with two simpler signals. The harmful label is populated at a similar rate in all three families (positive rate 0.072 in the passive family, 0.056 in the linked family and 0.053 in the action-revealed family), so the question is well posed everywhere; what differs is whether any signal rises above chance. In the action-revealed family the consistency residual and the predictive-variance baseline both rise above chance, but the ordering is not the one we proposed: the consistency residual reached AUROC 0.617, below the predictive-variance baseline at 0.666 but above the posterior-entropy baseline at 0.478, which is at chance or worse. In the linked family the same ordering holds (0.628 against 0.665 and 0.497). In the passive family, where the available gap is close to zero, the consistency residual is close to chance (0.519) and the two baselines straddle chance in opposite directions (0.575 and 0.432), so no signal there is reliably informative. A signal computed from the agent's own model therefore does contain information about when adaptation is harmful, but it does not outperform a simpler predictive-variance signal, and we report it as a negative result rather than as a contribution.

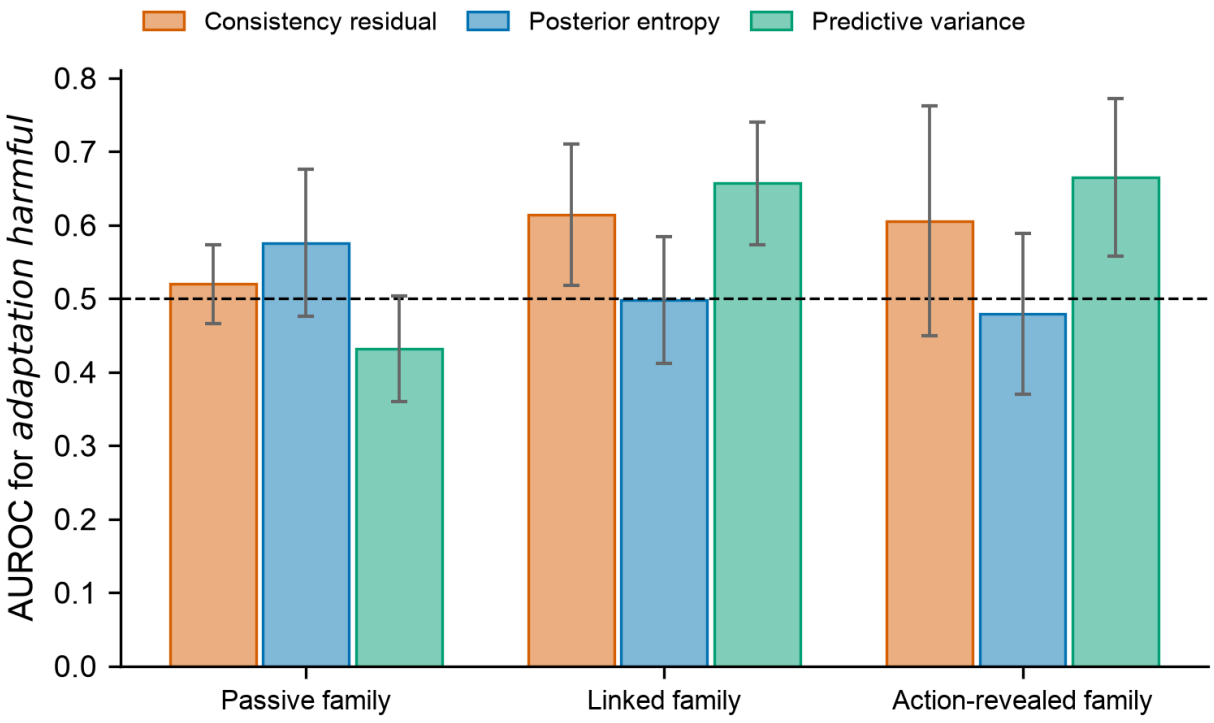



*Fig. 3.* A model-derived gate signal carries information but does not beat a simpler baseline. Discriminative performance of three label-free signals on the adaptation_harmful label (definition in electronic supplementary material, S6 Appendix), for the three families in a single panel. Each signal is evaluated on the five non-stationary drift kinds at strength 1.0, so each family contributes up to 125 cells (five drift kinds, 25 seeds); cells in which a signal is undefined are dropped, which leaves 121 in the action-revealed family. The proposed consistency residual attains AUROC 0.617 in the action-revealed family, 0.628 in the linked family and 0.519 in the passive family; it is below the predictive-variance baseline in the two informative families (0.666 and 0.665) but above the posterior-entropy

baseline (0.478 and 0.497). The AUROC values are means over cells; the error bar is the dispersion across those cells, and bootstrap 95% intervals computed within each seed and averaged across seeds are reported in S4 Table (`supplementary/s4_table_gating.md`). The positive rate of the label is 0.053 in the action-revealed family, 0.056 in the linked family and 0.072 in the passive family.

### 2.4 The adaptation advantage depends on the type of drift

The value of adaptation should also depend on how the regime moves (Fig 4). We separated stationary episodes from gradual and abrupt changes to transmission and reporting, a behavioural channel that is active only when the agent intervenes, and an abrupt loss of intervention efficacy, at an in-distribution and an out-of-distribution drift strength.

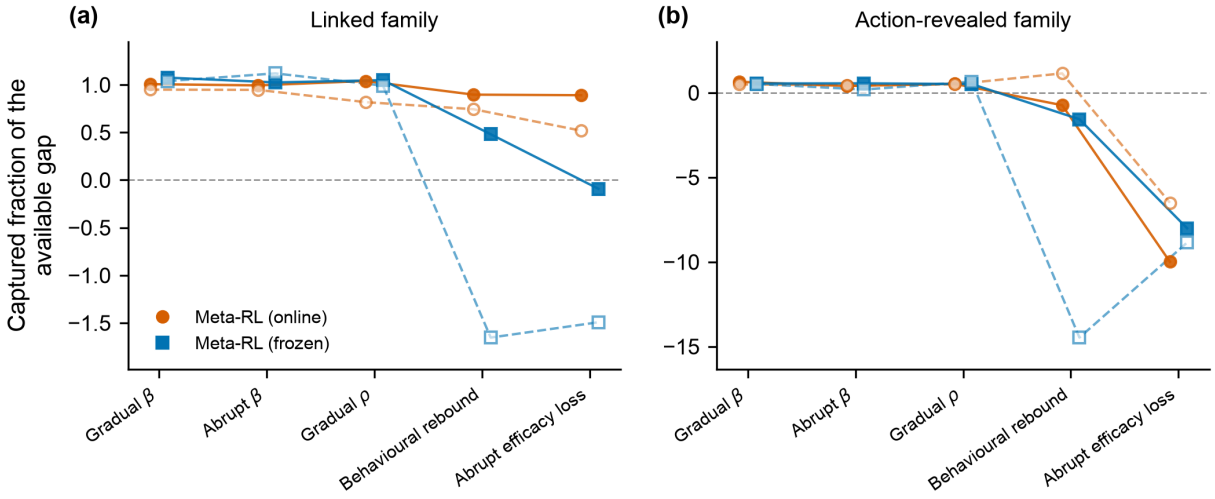



*Fig. 4.* The adaptation advantage depends on both drift type and identifiability. Full-episode captured fraction by drift kind for the online agent and the frozen ablation, at in-distribution (strength 1.0) and out-of-distribution (strength 2.0) drift, in the two families where that fraction is defined. Two panels: (a) linked, (b) action-revealed. The passive family is omitted, because its available gap is close to zero and the captured fraction is therefore not reported there (Section 3.1). Each panel shows the five non-stationary drift kinds (gradual_beta, abrupt_beta, gradual_rho, endogenous_beh, abrupt_eps; definitions in Section 4.3). For the online agent in panel (b) at strength 1.0 the captured fraction means are 0.65, 0.40, 0.53, -0.73 and -9.99; panel (a) uses the same two series, and its values, together with the frozen ablation in both panels, are provided as Source Data, including the per-kind standard deviations. Markers show means over the 25 training seeds; filled markers are strength 1.0 and open markers strength 2.0, and the panels use independent vertical scales. The behavioural and efficacy channels (endogenous_beh, abrupt_eps), which act only through the intervention, are where the two agents separate most in both panels. The abrupt_eps values in panel (b) are unstable for both agents, because several of its cells draw a small available gap and the captured ratio is then dominated by the denominator (Section 3.1).

### 2.5 Passive identifiability predicts where the regime's information has value

The boundary condition should be predictable before any meta-training is run. We tested whether the passive identifiability index, computed from passive trajectories alone, predicts where the regime's information has value. The index should be high wherever the parameters are passively readable, that is

in the passive and linked families, and close to chance in the action-revealed family; the prediction is that a high index coincides with a small value of adaptation and a near-chance index with a large one. The index separates the families as designed: its out-of-sample R2 is 0.591 in the linked family against 0.158 in the action-revealed family (S6 Appendix). The online-minus-frozen contrast, however, is small and straddles zero in both families (Section 2.2), so it does not by itself trace the predicted pattern; the between-family separation appears in the oracle-minus-online contrast (Section 2.2) and in the index itself. The index therefore functions as a design-time check on whether the regime's information is recoverable, which is the quantity that the boundary condition turns on.

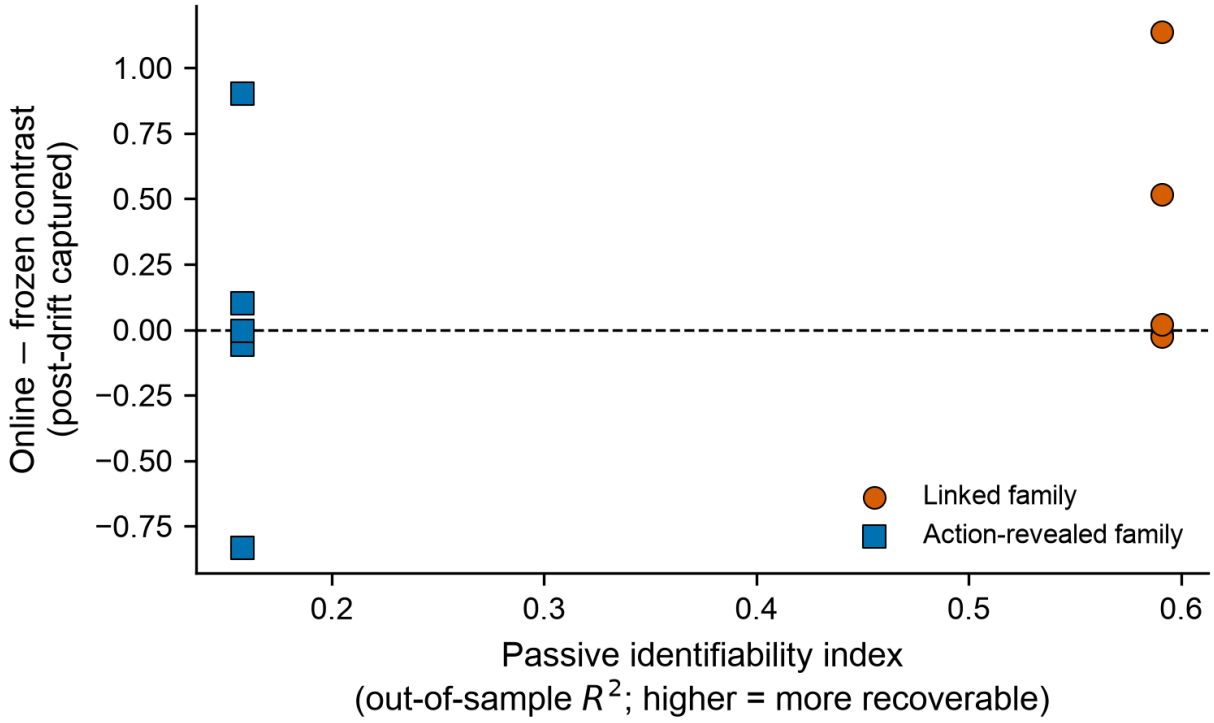



*Fig. 5.* Passive identifiability and the per-kind value of continuous re-inference. Each point is one drift kind within one family: the five non-stationary kinds in the linked family and the five in the action-revealed family, at drift strength 1.0. The horizontal axis is the out-of-sample R2 of the passive identifiability index for that family (0.591 in the linked family, 0.158 in the action-revealed family). It is constant within a family by construction, so the panel compares two index levels rather than tracing a continuous dose-response. The vertical axis is the paired post-drift difference in captured fraction between the online and frozen agents for that drift kind. In the linked family the five values are -0.02, -0.03, +0.02, +0.52 and +1.14; in the action-revealed family they are +0.10, -0.06, -0.00, +0.90 and -0.83, in both cases in the order gradual_beta, abrupt_beta, gradual_rho, endogenous_beh and abrupt_eps. Both sets straddle zero: the per-kind advantage of continuous re-inference over a frozen estimate does not separate the families, and the between-family separation appears instead in the oracle-minus-online contrast and in the index itself (Sections 2.2 and 2.5). The action-revealed abrupt_eps point rests on cells whose available gap is small, so its captured value is unstable (Section 3.1). The passive family is not shown, because its captured fraction is not reported there (Section 3.1). The per-family index values are reported in Section 2.5.

## 3 Discussion

We set out to test a specific claim about adaptive decision-making for epidemic intervention: that the benefit of explicit online regime inference depends on whether the parameters that matter for the decision can be recovered from passive observation, and not merely on whether a difference between interventions exists. The simulation supports that claim and rules out two simpler ones. Of the two predictions set out in the Introduction, the second held and the first did not.

Continuous re-inference gave no reliable advantage over an estimate frozen at the start of the episode in either family (+0.32 in the linked family and +0.02 in the action-revealed family, with across-seed standard deviations of 0.47 and 2.96). A drifting environment therefore does not, by itself, make it worthwhile to re-infer that drift continuously. The extra value of knowing the true regime was a different matter: +0.06 in the linked family against +2.03 in the action-revealed family (medians -0.06 and +0.22, with 9 and 17 of 25 seeds positive, and the action-revealed value larger in 18 of 25 seeds).

The mechanism accounts for both halves. Two tasks that differ only in the action channel produce identical observations until the agent acts, so an agent that never intervenes cannot tell them apart, and it is there that knowing the true regime was worth something. Where the regime is readable, the online agent already recovers essentially all of it. The expectation that a drifting environment rewards adaptation is not what the data show. The online agent beat the non-adapting baseline in the two families where the captured fraction is defined, but that baseline is weak in this environment (Section 3.1), and the comparison does not isolate adaptation. In the passive family the intervention parameters are fixed and the available gap is close to zero (2.53 under the experimental protocol), so that family serves only as a reference in which there is little to gain.

The mechanism is identifiability, and the three-family design is what isolates it. All three families face the same drift kinds and strengths, so a "more drift, more adaptation" account would predict similar gains throughout; what changes across families is whether the drifted quantities are observable without an action. The linked and action-revealed families are the informative contrast, because they are matched on the size of the available gap and differ only in whether that gap can be read from passive observation; the passive family is a reference in which the gap is near zero. Separating the existence of a decision-relevant difference from its identifiability is the methodological contribution of this study. The distinction itself is the same one that classical dual control draws between acting to control and acting to learn [18,19], specialised to a setting where the exploratory action is a public-health intervention and therefore carries a population-level cost. It is also close in spirit to decision-focused learning, which holds that a model should be judged by the decisions it induces rather than by the accuracy of its predictions [23,24]; here the decision-relevant quantity is whether a task is worth intervening on at all, and in the action-revealed family that quantity is not the family of quantities a passive predictive model can estimate.

Two measurements follow, of unequal value. The first is the passive identifiability index, which can be computed from a batch of passive trajectories before any meta-training. It is the cleanest construct in this study: measured out of sample it separates the families in the way the design intends, at 0.688 against a majority-class baseline of 0.484 in the linked family and 0.484 against 0.480 in the action-

revealed family, with the continuous variant agreeing (0.591 against 0.158). It is a design-time diagnostic: it tells a modeller, before any meta-training, whether the problem is one in which an adaptive meta-agent can be expected to help, or one in which a well-chosen fixed policy will do as well (Fig 5). The second is the gate signal, and here the result is negative. The consistency residual, the normalised one-step decoder error under the inferred context, does carry information about which adaptation steps are harmful, but a plain predictive-variance signal is at least as good in both families (Fig 3). We report this because it constrains the practical case for self-monitoring adaptation: an agent can compute a signal that is informative without regime labels, but the simplest such signal matches the model-derived one, so there is no evidence here that the agent's own context model is needed in order to know when to stop trusting it. Both quantities are computed only from observable quantities, which is a requirement rather than a convenience: in an epidemic the regime is never observed.

Our results are also consistent with the observation that, in the action-revealed family, most of the exploitable payoff comes from a coarse decision, whether to intervene at all, rather than from fine state-dependent modulation. When we bounded the achievable payoff by brute force on the action-revealed, stationary tasks (S6 Appendix), the best single constant action returned -164.37, the per-task-best constant action returned -134.65, and a simple state-dependent threshold rule returned -130.56. The gap between the best single constant and the per-task-best constant is therefore almost the whole of the available payoff, and it is realised by identifying the task, not by reacting to the observed state (S6 Appendix). This is why an agent that fails to resolve the action-channel ambiguity cannot capture the gap even with a perfect reactive component, and why the boundary condition we report is about identifiability rather than about policy expressiveness.

The failure mode of the non-adaptive baseline is the one this design predicts, and it is measurable. Trained by domain randomisation with no task conditioning, the baseline converges to a policy that is nearly constant in the action. On the tasks whose per-task optimal action is the no-intervention level, which are 60.1% of the linked tasks, it plays the strongest intervention level on roughly three quarters of its steps and captures only 1.58 of the available gap. On the tasks whose optimal action is the intermediate level it captures 5.61 (880 tasks). Two interventions that should help a weak learner barely move it: doubling the pre-training budget to 2000 iterations changed its mean post-drift captured fraction by +0.13 (standard deviation 0.43 across seeds), and giving it a 12-step observation history changed it by +0.08 (standard deviation 0.60). The deficit is therefore not explained by under-training or by missing memory, and it is concentrated exactly where a per-task decision is required.

That is direct evidence for the claim of this paper. What the meta-agents have over this baseline is the conditioning of the action on the task, not a better learner, so a method whose inductive bias is memoryless domain randomisation forfeits the gap on precisely the tasks where conditioning matters. The 60.1% share is distinct from the 0.38 share of linked tasks on which the behavioural rebound dominates (S6 Appendix); even where an intervention lowers transmission, the cost of intervening can leave no intervention as the better constant action, which is why the two counts differ. The diagnostic is tabulated in S5 Table.

A related body of work brings meta-RL into contact with problems that are close to ours but not identical. Transfer RL for time-varying, partially observable epidemic parameters is the nearest direct competitor and shares the premise that the parameters that set the best action are unknown and move [9]; Bayesian contextual bandits for public-health resource allocation share the concern with unobserved context but treat the problem as allocation rather than as repeated control of a dynamical system [8]. Meta-RL in non-stationary environments [16] and self-supervised task representations [17] provide the machinery we build on. What is different here is that we do not propose a new algorithm; we hold the algorithm fixed and vary the identifiability of the environment, which lets us isolate the boundary condition that determines whether the family of methods is applicable at all.

### 3.1 Limitations

Eight limitations bound the interpretation of these results.

First, the results are from simulation, not from field deployment. The environment is a seasonal SEIRS-type transmission model with a reporting layer, and although a variety of real surveillance series were used to calibrate the drift magnitude (Fig 1), the intervention dynamics themselves are a simplification of any real response. A field trial of an adaptive intervention policy would face partial compliance, heterogeneity across subpopulations, and political constraints that the simulator does not represent.

Second, the agent observes only reported case counts and calendar information. Real surveillance systems carry more: hospitalisations and deaths, genomic sequencing to identify variants, seroprevalence, mobility and contact surveys, and policy-stringency records. Any of these could make part of the action channel passively identifiable and would therefore weaken the boundary condition reported here. Our claim is about what can be inferred from a minimal, widely available observation stream, not about the ceiling of what is inferable in principle.

Third, the drift magnitudes were calibrated to real data, but the mechanisms that generate the drift are simplified. The simulated in-distribution drift corresponds closely to the median structural change detected in real influenza-like-illness surveillance, and even the out-of-distribution condition remains inside the range of changes actually observed, so the drift is not an artificial extreme. The channels, however (a shift in baseline transmissibility, in the reporting rate, in behaviour, or in efficacy), are stylised versions of events that in reality co-occur and interact.

Fourth, meta-RL training on this environment shows seed variance, and with 25 training seeds the across-seed distributions of the paired contrasts are heavy-tailed. The tail is driven by a small number of cells whose available gap is small, where the captured fraction divides by a near-zero denominator, the same pathology that makes the fraction unreportable in the passive family. In the action-revealed family the oracle-minus-online contrast has a median of +0.22 and an interquartile range of -0.06 to +1.18, with 17 of 25 seeds positive, while its mean of +2.03 is inflated by the most extreme of those cells. Diagnostic runs at different training budgets (S1 Text, S1 Fig) moved the captured fraction in both directions, so the absolute performance numbers should be read as indicative of the boundary condition rather than as stable benchmarks, and any single-run headline number should be treated with caution. The boundary condition itself, the across-family ordering of the oracle-minus-online contrast, was

reproduced in 18 of 25 seeds in captured units and in 17 of 25 in raw post-drift return, and it is the property we intend the paper to support.

Fifth, the linked family is a deliberately simplified model of the situation it represents. Its premise is that the two parameters governing the effect of intervention are affine functions of the baseline transmission rate, so that they are exactly recoverable from the passive case trajectory. In a real epidemic the link between observable circulation and the response to an intervention is noisy, partial and confounded with reporting changes, so the linked family should be read as an idealised case in which identifiability holds by construction, not as a description of how identifiability arises in practice. Its role is to make the boundary condition falsifiable: it supplies a large decision-relevant difference that is nevertheless passively visible, so that a method's failure to adapt there cannot be attributed to the absence of a gap.

Sixth, the non-adaptive baseline is weak in this environment, and the weakness is diagnosed rather than assumed. On the post-drift window it captures 4.95 in the linked family, against 5.12 for the transfer baseline and 4.82 for the retraining baseline, with across-seed standard deviations near 8, so the three are close and the comparison is dominated by seed variance. The diagnostic in S5 Table traces the weakness to the inductive bias of domain randomisation, not to under-training or to missing memory. The baseline converges to a policy that is nearly constant in the action, so on the 60.1% of linked tasks whose per-task optimal action is no intervention (880 tasks) it captures only 1.58 of the available gap, against 5.61 on the tasks whose optimal action is the intermediate level. Neither doubling the pre-training budget (+0.13) nor adding a 12-step observation history (+0.08) recovers the deficit. We therefore make no claim that adaptation is better than non-adaptation: an online-minus-non-adapting contrast would be dominated by the weakness of that particular baseline, which is why every comparison in this paper is drawn against the frozen and oracle agents instead.

Seventh, the captured fraction is unreliable in the passive family, and we do not report it there. The fraction divides by the per-task available gap, and in the passive family that gap is close to zero, so the ratio is numerically unstable and its across-seed spread is dominated by the small denominator rather than by agent behaviour. For this reason the passive family is reported as raw return, it is never plotted or compared as a captured fraction, and the family-level statement we make about it is limited to the observation that its available gap is small (2.53 under the experimental protocol, against 31.12 and 25.33 for the other two families). The same instability is not confined to the passive family: eight cells among the linked and action-revealed drift conditions draw an available gap below 5, and those cells dominate the across-seed tails of the captured contrasts; the medians and per-seed sign counts reported alongside the means in S3 Table are what make the contrasts readable despite that tail.

Eighth, the results are not bit-reproducible across thread counts: a control re-run of the non-adaptive baseline with two threads rather than four correlates with the reported run at 0.67 across the 55 linked cell-seed observations and differs in the mean by 0.46 captured units. The qualitative ordering of the agents is unchanged, but exact values depend on the thread count, which is recorded in the provenance block of every result file.

### 3.2 Implications

These results bear on how adaptive artificial intelligence for epidemic response is framed. It is easy to assume that an environment which changes requires an agent that adapts, and that an agent which adapts will improve decisions. The simulation gives that assumption two preconditions: a decision-relevant difference must exist, and it must not already be readable from passive observation.

Where the parameters that determine the best action can be read from passive surveillance, a simpler state-dependent strategy is adequate, even when the difference between interventions is large, because inference already recovers what an oracle would know. Where those parameters can only be learned by acting, the value of knowing them is real, but realising that value through online inference is the hard part, and in this environment continuous re-inference did not achieve it. The identifiability index says which regime one is in before any training is run.

The same logic bounds how long the problem lasts. As surveillance systems broaden and include more of the signals listed above, more of the action-channel ambiguity becomes resolvable without intervening, and the class of problems for which adaptive meta-control is warranted shrinks. Deciding which regime a real outbreak presents is the first question to ask before committing to an adaptive policy.

## 4 Materials and Methods

**Environment.** The environment is a batched, pure-NumPy seasonal SEIRS-type transmission model with a negative-binomial reporting layer, advanced in weekly steps. Weekly transmissibility is modulated seasonally [2,3] and by the intervention: a direct efficacy term lowers transmission and a behavioural rebound term raises contacts, so intervening is not always beneficial. The agent chooses one of three intervention levels per week; the reward combines the infected fraction (reference scale 0.02) with a per-step intervention cost, which makes the reward non-monotone in intensity. The agent observes a four-dimensional vector (scaled log case count, its weekly change, the seasonal cosine and the previous action); the latent regime is never observed. Reported cases are generated from the exposed fraction with reporting rate rho, which is itself part of the regime, reflecting the fact that the mapping from true infections to reported cases is uncertain [27,28]. The model is a deliberately compact relative of seasonal transmission models [25] and of individual-based simulators such as Covasim [26]; the decision problem, not the mechanistic detail, is the object of study. The full dynamical specification, the observation and reward definitions and the reporting-layer parameters are given in electronic supplementary material, S6 Appendix.

**Environment families.** Three families share the decision space and the observation process and differ only in how the two action-channel parameters (direct efficacy eps, behavioural rebound beh) are set. In the passive family they are fixed, so the regime is identifiable without intervening but there is little to gain (available gap 0.92). In the linked family they vary across tasks but are affine functions of the passively visible baseline transmission rate, so a large decision-relevant difference exists and is exactly readable from passive trajectories (available gap 24.99). In the action-revealed family they vary independently and enter the dynamics only through the action term, so two tasks that differ in them produce identical observations until the agent intervenes (available gap 25.26); this property is verified

rather than assumed, by an environment-validation check that confirms zero passive-observation difference for action-channel-identical tasks. The linked and action-revealed families are matched on the size of the available gap, so the contrast between them isolates identifiability from how much is at stake. Full family definitions and the drift conditions are given in electronic supplementary material, S6 Appendix.

**Agents and algorithm.** Four adaptive agents share one architecture: an online agent that re-infers the latent regime from the most recent 12 transitions at every step, a frozen ablation that infers once and freezes the estimate, an oracle given the true regime, and three non-meta baselines (domain-randomised non-adaptive, transfer with fine-tuning, per-task retraining). The meta-agents follow the PEARL design [13]: a permutation-invariant context encoder maps recent transitions to a diagonal Gaussian posterior over a latent context, and the policy and value functions are conditioned on the observation and the posterior mean. The policy is trained with PPO [30]. Full hyper-parameters and the decoder-based auxiliary objective are given in electronic supplementary material, S6 Appendix.

**Protocol, data and validation.** Meta-agents are trained on the stationary task distribution of each family and evaluated on drifting episodes (five non-stationary drift kinds at two strengths plus stationary, 25 training seeds). The primary contrasts are computed on the post-drift window, in which adaptation is visible: the online-minus-frozen difference and the oracle-minus-online difference, both in captured-fraction units (the share of the per-task available gap an agent closes). Drift magnitude was calibrated against real respiratory surveillance: the in-distribution condition corresponds to roughly the median structural change detected in US influenza-like-illness and WHO FluNet series [20,21,22], and the out-of-distribution condition remains inside the observed range. Intervention policies in real response data are recorded as graded stringency indices [29]; the three discrete levels here are an abstraction of that continuum. The environment passed a 23-check validation suite, including the zero passive-observation difference check (S1 Table). All experiments ran on CPU only. The complete protocol, data sources and validation checks are given in electronic supplementary material, S6 Appendix.

## Acknowledgments

The author thanks the maintainers of the public surveillance resources used for calibration ( Delphi Epidata, WHO FluNet, and the open COVID-19 case series) for making the data available.

## References

1. Kraemer MUG, Tsui JLH, Chang SY, Lytras S, Khurana MP, Vanderslott S, et al. Artificial intelligence for modelling infectious disease epidemics. Nature. 2025;638(8051):623-635. doi:10.1038/s41586-024-08564-w

2. Shaman J, Kohn M. Absolute humidity modulates influenza survival, transmission, and seasonality. Proc Natl Acad Sci U S A. 2009;106(9):3243-3248. doi:10.1073/pnas.0806852106

3. Dushoff J, Plotkin JB, Levin SA, Earn DJD. Dynamical resonance can account for seasonality of influenza epidemics. Proc Natl Acad Sci U S A. 2004;101(48):16915-16916. doi:10.1073/pnas.0407293101

4. Libin PJK, Moonens A, Verstraeten T, Perez-Sanjines F, Hens N, Lemey P, et al. Deep reinforcement learning for large-scale epidemic control. In: Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track (ECML PKDD 2020), Lecture Notes in Computer Science, vol 12461. Cham: Springer; 2021. p. 155-170. doi:10.1007/978-3-030-67670-4_10

5. Kwak GH, Ling L, Hui P. Deep reinforcement learning approaches for global public health strategies for COVID-19 pandemic. PLoS One. 2021;16(5):e0251550. doi:10.1371/journal.pone.0251550

6. Rusu AC, Farrahi K, Niranjan M. EpiCURB: learning to derive epidemic control policies. IEEE Pervasive Comput. 2024;23(1):57-62. doi:10.1109/MPRV.2023.3329546

7. Feng T, Xia T, Fan X, Wang H, Zong Z, Li Y. Precise mobility intervention for epidemic control using unobservable information via deep reinforcement learning. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM; 2022. p. 2882-2892. doi:10.1145/3534678.3539195

8. Liang B, Xu L, Taneja A, Tambe M, Janson L. Context in public health for underserved communities: a Bayesian approach to online restless bandits. Proc AAAI Conf Artif Intell. 2025;39(27):28195-28203. doi:10.1609/aaai.v39i27.35039

9. Du M, Yu H, Kong N. Transfer reinforcement learning for mixed observability Markov decision processes with time-varying interval-valued parameters and its application in pandemic control. INFORMS J Comput. 2025;37(2):315-337. doi:10.1287/ijoc.2022.0236

10. Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning. PMLR; 2017. p. 1126-1135.

11. Duan Y, Schulman J, Chen X, Bartlett PL, Sutskever I, Abbeel P. RL^2: Fast reinforcement learning via slow reinforcement learning. arXiv:1611.02779; 2016.

12. Wang JX, Kurth-Nelson Z, Tirumala D, Soyer H, Leibo JZ, Munos R, et al. Learning to reinforcement learn. arXiv:1611.05763; 2016.

13. Rakelly K, Zhou A, Finn C, Levine S, Quillen D. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In: Proceedings of the 36th International Conference on Machine Learning. PMLR; 2019. p. 5331-5340.

14. Zintgraf L, Shiarlis K, Igl M, Schulze S, Gal Y, Hofmann K, et al. VariBAD: A very good method for Bayes-adaptive deep RL via meta-learning. In: International Conference on Learning Representations (ICLR); 2020.

15. Beck J, Vuorio R, Liu EZ, Xiong Z, Zintgraf L, Finn C, et al. A tutorial on meta-reinforcement learning. Found Trends Mach Learn. 2025;18(2-3):224-384. doi:10.1561/2200000080

16. Bing Z, Lerch D, Huang K, Knoll A. Meta-reinforcement learning in non-stationary and dynamic environments. IEEE Trans Pattern Anal Mach Intell. 2023;45(3):3476-3491. doi:10.1109/TPAMI.2022.3185549

17. Wang M, Bing Z, Yao X, Wang S, Huang K, Su H, et al. Meta-reinforcement learning based on self-supervised task representation learning. Proc AAAI Conf Artif Intell. 2023;37(8):10157-10165. doi:10.1609/aaai.v37i8.26210

18. Filatov NM, Unbehauen H. Survey of adaptive dual control methods. IEE Proc Control Theory Appl. 2000;147(1):118-128. doi:10.1049/ip-cta:20000107

19. Chen WH. Perspective view of autonomous control in unknown environment: dual control for exploitation and exploration vs reinforcement learning. Neurocomputing. 2022;497:50-63. doi:10.1016/j.neucom.2022.04.131

20. Reinhart A, Brooks L, Jahja M, Rumack A, Tang J, Agrawal S, et al. An open repository of real-time COVID-19 indicators. Proc Natl Acad Sci U S A. 2021;118(51):e2111452118. doi:10.1073/pnas.2111452118

21. Flahault A. FluNet as a tool for global monitoring of influenza on the Web. JAMA. 1998;280(15):1330. doi:10.1001/jama.280.15.1330

22. Newman LP, Bhat N, Fleming JA, Neuzil KM. Global influenza seasonality to inform country-level vaccine programs: an analysis of WHO FluNet influenza surveillance data between 2011 and 2016. PLoS One. 2018;13(2):e0193263. doi:10.1371/journal.pone.0193263

23. Elmachtoub AN, Grigas P. Smart "predict, then optimize". Manage Sci. 2022;68(1):9-26. doi:10.1287/mnsc.2020.3922

24. Mandi J, Kotary J, Berden S, Mulamba M, Bucarey V, Guns T, et al. Decision-focused learning: foundations, state of the art, benchmark and future opportunities. J Artif Intell Res. 2024;80:1623-1701. doi:10.1613/jair.1.15320

25. Weber A, Weber M, Milligan P. Modeling epidemics caused by respiratory syncytial virus (RSV). Math Biosci. 2001;172(2):95-113. doi:10.1016/S0025-5564(01)00066-9

26. Kerr CC, Stuart RM, Mistry D, Abeysuriya RG, Rosenfeld K, Hart GR, et al. Covasim: an agent-based model of COVID-19 dynamics and interventions. PLoS Comput Biol. 2021;17(7):e1009149. doi:10.1371/journal.pcbi.1009149

27. Gibbons CL, Mangen MJJ, Plass D, Havelaar AH, Brooke RJ, Kramarz P, et al. Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. BMC Public Health. 2014;14(1):147. doi:10.1186/1471-2458-14-147

28. Reed C, Chaves SS, Daily Kirley P, Emerson R, Aragon D, Hancock EB, et al. Estimating influenza disease burden from population-based surveillance data in the United States. PLoS One. 2015;10(3):e0118369. doi:10.1371/journal.pone.0118369

29. Hale T, Angrist N, Goldszmidt R, Kira B, Petherick A, Phillips T, et al. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). Nat Hum Behav. 2021;5(4):529-538. doi:10.1038/s41562-021-01079-8

30. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. arXiv:1707.06347; 2017.

## Electronic supplementary material

S1 Text. Diagnostic ladder and training-budget analysis (single-seed diagnostics; file supplementary/s1_text_diagnostics.md). S1 Fig. Training-budget diagnostic ladder (file figures/figS1_diagnostic_ladder.png). S1 Table. Environment validation checklist (23 mechanistic checks; file supplementary/s1_table_env_checks.md). S2 Table. Drift calibration summary (real quantiles versus simulated strengths; file supplementary/s2_table_drift_calibration.md). S3 Table. Return, captured fraction, regret and adaptation speed by family and agent, with paired per-seed contrasts (file supplementary/s3_table_performance.md). S4 Table. Gating AUROC by family and signal, with bootstrap intervals and label positive rates (file supplementary/s4_table_gating.md). S5 Table. Diagnosis of the non-adaptive baseline (file supplementary/s5_table_baseline.md). S6 Appendix. Full model, agents and experimental protocol. S1 Data. Source data and code (public repository, see Data availability).

## Declarations

**Generative AI statement.** During the preparation of this work the author used Kimi K3 (Moonshot AI) for code development and language editing. The author reviewed and edited all AI-generated content and takes full responsibility for the integrity and accuracy of the publication.

**Competing interests.** The author has declared that no competing interests exist.

**Funding.** The author received no specific funding for this work.

**Data availability.** The real surveillance series analysed in this study are publicly available: US influenza-like-illness activity from the Delphi Epidata FluView stream and the WHO FluNet positivity stream, as cited in Materials and Methods. The simulated data, the environment, the agents, the gating diagnostics and the identifiability index are generated by the code in S1 Data; the parameters that fully specify the simulation are given in Materials and Methods. The complete codebase and the figure source data are also publicly available at https://github.com/GustaveKateb0709/online-meta-rl-epidemic-drift.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigS1.tif
- Supplementary.zip