SAFE EXECUTIONS CAN HIDE POLICY FAILURE: A PROPOSAL-LEVEL AUDIT OF HYBRID CONTROL IN SYNTHETIC INTERVIEWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Safe executions can conceal unsafe proposals or a policy that avoids useful actions. We audit these failure modes in a synthetic structured-interview benchmark using joint proposal, execution, intervention and utility traces. A common hybrid-action interface supports TD3-style and PPO controllers with memory, reliability conditioning, nuisance consistency and proposal-cost optimization. Across 60 fitted policies and 23,520 evaluation episodes, both full controllers avoid follow-ups in the nominal panel, while a transparent rule achieves higher utility. Component-stack effects reverse direction between algorithms. An auxiliary control isolates exploration-induced constraint pressure: training exploration raises an always-proceed controller's mean discounted proposal cost to 0.696, above its budget of 0.5, despite zero deployed proposal cost. Released checkpoints, turn traces and executable audits connect every outcome table to its source, with explicit denominators and paired subject-level uncertainty. The findings motivate evaluating executed safety jointly with proposal safety and useful activity, and distinguishing exploratory behavior costs from target-policy costs. The evidence concerns a controlled synthetic setting; real-audio transfer and clinical effectiveness remain untested.

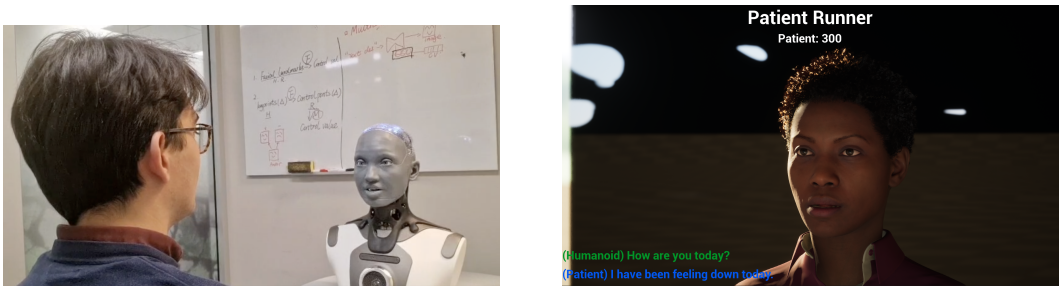


Figure 1: Motivating platform images: Ameca interaction (left) and the original patient-renderer example (right). Reported experiments use InterviewLite, not the pictured robot or renderer.

1 INTRODUCTION

An adaptive interviewer faces two coupled decisions: when to speak and what kind of bounded follow-up, if any, to select. An action can be inappropriate because of its timing, its content category, or its dependence on unreliable observations. A policy optimized only for interaction reward may learn behavior that is corrected by a downstream rule system. Reporting the corrected behavior alone does not identify what the policy learned. This separates policy optimization from intervention, two established routes to safe reinforcement learning (García & Fernández, 2015; Alshiekh et al., 2018).

The motivating application is a structured, clinician-supervised screening interview. We use a fixed questionnaire backbone and a limited number of follow-ups in an independently executable synthetic study. No diagnostic labels, clinical language model, audio encoder, rendered patient, or human

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

response is present in the experiment. "Patient" in the implementation names a synthetic subject identifier and the statistical clustering unit; it does not denote a study participant.

Our contributions are: (i) a joint proposal–execution–utility audit that distinguishes shield intervention from policy avoidance; (ii) a paired algorithm-by-component study showing that stack effects depend on the learning backbone; and (iii) an exploration-only control that exposes a mismatch between deployed-policy cost and the costs used for multiplier updates. A fully specified hybrid interface and released checkpoints, traces and reconstruction checks make each finding inspectable. These contributions concern evaluation and diagnosis; they do not establish a superior interviewing algorithm.

Three favorable-looking metrics can therefore mislead: perfect executed safety may reflect a shield, low proposal cost may reflect avoidance, and complete item coverage may be guaranteed by the environment. We report the associated utility and constraints explicitly. The contribution is an auditable failure analysis of these effects.

2 Related Work and Positioning

Social interaction and sensing. Long-term social-robot interaction motivates evaluating more than immediate task completion (Breazeal et al., 2016; Leite et al., 2013). A structured interviewer additionally has a content obligation: a fluent conversation can still omit a required item, and completing every item can still produce an unhelpful interaction. Speech-based depression and risk assessment has its own measurement challenges (Cummins et al., 2015). We therefore separate interviewing behavior from inference about clinical state. InterviewLite measures neither symptom severity nor diagnostic accuracy, and its symbolic distress flag should not be interpreted as a validated speech-derived risk estimate.

The Distress Analysis Interview Corpus provides a relevant setting for studying interviews between people and human or computer interviewers (Gratch et al., 2014). The AVEC 2019 challenge provides the corresponding extended-corpus benchmark context (Ringeval et al., 2019). Neither corpus is used in our synthetic experiment. Virtual interviewing in SimSensei and communication coaching in MACH illustrate related interaction settings (DeVault et al., 2014; Hoque et al., 2013). Corpus-based recognition and online intervention have different information requirements: a model that consumes an entire response can evaluate that response, but it cannot justify a timing action taken halfway through it. A reproducible transfer study must state exactly which recording channels and time prefixes produce each controller input.

Speech recognition (Radford et al., 2023), speaker verification (Desplanques et al., 2020), facial analysis (Baltrušaitis et al., 2018) and pose estimation (Cao et al., 2017) offer distinct observation channels; none alone establishes readiness to yield the floor. EEG-based multimodal mental-health evaluation (Hu et al., 2024) and interpretable speech-level depression modeling (Deng et al., 2023) address different targets. InterviewLite's generated scalar quality scores provide no robustness evidence about these external encoders.

Dialogue policies and turn timing. Dialogue-policy surveys distinguish policy learning from broader language-model development (Kwan et al., 2023; Wang et al., 2023). Prior work studies language-valued actions (He et al., 2016), interactive reward acquisition (Su et al., 2016) and actor–critic task completion (Peng et al., 2018). Our categorical action space makes proposal accounting inspectable, but category selection does not establish that a realized utterance is appropriate or understandable.

Turn-taking depends on predictive coordination rather than silence duration alone (Skantze, 2021; Levinson & Torreira, 2015). Voice Activity Projection predicts future voice activity and supports turn-taking event modeling (Ekstedt & Skantze, 2022). Its application with TurnGPT in human–robot conversation supplies a concrete comparison design (Skantze & Irfan, 2025). It is a more direct timing comparator than a generic dialogue return alone. A fair comparison would expose both systems to the same causal audio prefix and convert each output into a declared decision rule. It would report the trade-off between speaking too soon and waiting too long, including cases where a system elects not to speak. The synthetic scalar observations in this work cannot support such a

comparison; the reference implementation instead defines the action and logging interface to which a real-audio adapter can connect.

Learning backbones, simulators and transfer. TD3 and PPO supply different policy-optimization mechanisms (Fujimoto et al., 2018; Schulman et al., 2017). Their inclusion allows a limited test of whether a component contrast depends on the backbone. It does not establish coverage of all plausible algorithms. Offline approaches such as implicit Q-learning and conservative Q-learning, or a sequence-model policy such as Decision Transformer, address different training regimes (Kostrikov et al., 2021; Kumar et al., 2020; Chen et al., 2021). Comparing them would require a declared shared logged dataset and a leakage-free selection protocol. Generating new online episodes for one method while limiting another to a poor behavior dataset would confound algorithm and information access.

Domain randomization varies simulated training conditions (Tobin et al., 2017), while CARLA exemplifies an explicitly specified task simulator (Dosovitskiy et al., 2017); neither establishes interviewing fidelity. Work on robot communicative style and embodiment (Banck et al., 2025; Banna et al., 2025), and human-like game agents (Świechowski & Ślęzak, 2025), further distinguishes task scores from perceived interaction quality. Figure 1 illustrates the motivating platform, whereas the empirical contribution is a checkable control experiment with an explicit synthetic evaluation contract.

3 Problem and Action Contract

We model the controller as acting on a partially observed process (Kaelbling et al., 1998). At each step it receives observations o_t, selects a proposal a_t, and observes the transition resulting from executed action $\tilde{a}_t = G(o_t, a_t)$. The objective is

$$\max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{T-1} \gamma^t r_t\right] \quad \text{subject to} \quad \mathbb{E}_{\pi}\left[\sum_{t=0}^{T-1} \gamma^t c_t\right] \leq d. \tag{1}$$

The cost c_t labels an unsafe *proposal*, including one blocked by G. This prevents the shield from hiding proposal errors from the learning objective. Equation 1 is a target objective; the approximate neural optimization does not guarantee feasibility. Unlike constrained policy optimization (Achiam et al., 2017), we do not derive a trust-region feasibility bound.

The normalized continuous action $u_t \in [0, 1]^4$ maps to response latency $[0.1, 1.6]$ seconds, maximum wait $[0.2, 2.0]$ seconds, backchannel level $[0, 0.5]$, and interruption tolerance $[0, 0.2]$. In the current fixture only latency and maximum wait affect transitions. The other outputs preserve interface slots and are logged, but no result claims that they learn meaningful backchanneling or interruption control.

The discrete action $k_t \in \{0, \ldots, 6\}$ selects proceed, frequency clarification, temporal clarification, functional-impact probing, cluster clarification, safety clarification, or repair. These are symbolic categories, not generated utterances. A seven-way categorical head replaces an unspecified continuous gate followed by an external selector. At evaluation we execute the category with maximum probability. A proceed action advances one mandatory item; a probe consumes one of nine adaptive turns. All completed fixture episodes satisfy

$$T = 25 + A, \qquad 0 \leq A \leq 9. \tag{2}$$

The structural budget remains active in the shield-off condition. Thus "shield off" removes the safety rule, not the definition of the action budget.

4 Controller and Learning

4.1 Reliability and finite history

The observation has sixteen scalar entries: four cues, four quality scores, and eight ledger/action-summary entries. The ledger contains item progress, remaining budget, refusal, distress, ambiguity, consecutive-probe count, previous category and previous latency. The first two cues are noisy timing

measurements; the last two are nuisance variables. Safety flags are directly visible in this fixture. This makes the safety problem substantially easier than recognizing risk from speech.

Reliability conditioning multiplies each cue by its quality score and retains the quality vector. A linear layer with 32 outputs and a hyperbolic-tangent activation feeds a 32-unit GRU (Cho et al., 2014) over the most recent eight observations. Histories are left-zero-padded at episode start and reset between episodes. Recurrent control can aggregate incomplete observations (Hausknecht & Stone, 2015); our finite window is not a calibrated belief state. Each actor and critic has its own encoder. The no-memory ablation uses only the final encoded observation. The no-reliability ablation removes both multiplication and explicit quality inputs; missing cues can still be detectable through their values.

4.2 Hybrid TD3-style policy

We retain twin reward critics, target-policy smoothing, delayed actor updates and slowly updated target networks from TD3 (Fujimoto et al., 2018). The categorical extension is explicit and differs from vanilla continuous-action TD3. The actor produces $u_\theta(h)$ and categorical probabilities $p_\theta(k \mid h)$, where h is the history window. Critics take a continuous action and an exact one-hot category. For target evaluation,

$$u' = \mathrm{clip}_{[0,1]}\big(u_{\bar{\theta}}(h') + \epsilon\big), \qquad k' = \arg\max_k p_{\bar{\theta}}(k \mid h'), \tag{3}$$

$$y_r = r + \gamma(1 - \mathrm{done}) \min_j Q^r_{\bar{\phi}_j}(h', u', k'), \tag{4}$$

$$y_c = c + \gamma(1 - \mathrm{done}) \max_j Q^c_{\bar{\psi}_j}(h', u', k'). \tag{5}$$

We enumerate all seven one-hot categories in the actor loss, rather than training a critic on a fractional category that is never executed:

$$\mathcal{L}_\pi = -\mathbb{E}_h \sum_k p_\theta(k \mid h) \left[Q^r_{\phi_1}(h, u_\theta(h), k) - \lambda \max_j Q^c_{\psi_j}(h, u_\theta(h), k) \right] + \eta \mathcal{L}_{\mathrm{cf}}. \tag{6}$$

This smooth surrogate trains the categorical head; deployment uses its mode. The surrogate/mode mismatch and cost-critic approximation are limitations, so learned feasibility is measured rather than assumed.

We update the nonnegative multiplier from each newly collected episode,

$$\lambda \leftarrow \mathrm{clip}_{[0,20]} \left(\lambda + 0.02 \left[\sum_t \gamma^t c_t - 0.5 \right] \right), \qquad \gamma = 0.97. \tag{7}$$

The discounted proposal budget is $d = 0.5$ per training episode; it is not a percentage and does not imply a clinical safety threshold. The no-cost setting fixes $\lambda = 0$. We retain the critic architecture when disabling its contribution so that the ablation changes the objective rather than network capacity. PID Lagrangian updates address oscillatory multiplier dynamics (Stooke et al., 2020); they are a prospective comparator, not the update used in the reported matrix.

4.3 PPO and common consistency loss

The PPO controller (Schulman et al., 2017) uses the same encoder choices, a categorical head, and independent Beta distributions for bounded continuous actions. It uses sampled actions in training and distribution means plus the categorical mode in evaluation. Separate reward and cost value outputs form a penalized advantage. We use discounted full-episode returns, four optimization epochs per episode, a clipped likelihood ratio, and the same online multiplier update. This provides a second algorithm with matched observations, rewards, costs and component availability; equal episode budgets do not imply equal optimizer work or equal sample efficiency.

The consistency loss is shared:

$$\mathcal{L}_{\mathrm{cf}} = \tfrac{1}{4} \| u(h) - u(h^{\mathrm{cf}}) \|_2^2 + \mathrm{KL}\big(p(\cdot \mid h) \,\|\, p(\cdot \mid h^{\mathrm{cf}})\big). \tag{8}$$

Only the two known nuisance cue coordinates are perturbed, with clipped Gaussian noise of standard deviation 0.05; timing cues, quality, ledger and safety flags are preserved. We use $\eta = 0.1$. This is a nuisance-invariance regularizer. It does not demonstrate a causal counterfactual model of people, and it must not be applied indiscriminately to informative prosody or distress cues.

5 SYNTHETIC ENVIRONMENT AND MEASUREMENT

InterviewLite is an executable fixture with a fixed synthetic subject-speed parameter, a smoothly varying latent pause duration and noisy observations. It does not derive parameters from E-DAIC. Nominal training uses the safety shield. Evaluation includes nominal conditions, dropout of the first and third cue streams with zeroed quality, and a slow condition with a 0.45-second pause extension capped at 1.6 seconds. The stress settings are engineering probes, not models of demographic or clinical groups.

An ordinary probe is unsafe when refusal or distress is present, its budget is exhausted, two consecutive probes have occurred, or mean quality falls below 0.4. The optional shield replaces an unsafe probe with proceed. The actual application would require a validated escalation and handover mechanism; proceeding in this symbolic fixture is not a clinical response to crisis. A structurally exhausted budget also forces proceed with a separately logged reason.

The fixture emits at the minimum of the selected latency and maximum wait. Overlap is the positive part of latent pause duration minus emission time. A probe is useful only when its category matches a deterministic item-associated category, ambiguity is present, the item has not already been clarified, and execution is safe. This transparent construction tests the pipeline but is not a measure of clinical usefulness. Objective misspecification remains possible even when optimization succeeds (Amodei et al., 2016).

There is exactly one reward definition in the executable study:

$$r_t = 1\Delta C_t + 0.15U_t - 0.30O_t - 0.02D_t - 0.01P_t. \tag{9}$$

Here $\Delta C_t = 1/25$ for proceed and zero otherwise; U_t is useful clarification, O_t is overlap seconds, D_t is excess response delay, and P_t indicates an executed probe. We remove unsupported screening-sensitivity and diagnostic-consistency heads from the empirical claim. Every component is stored in the trace and the scalar reward is reconstructed from it. Coverage is always complete by construction and is not a learned-performance outcome.

5.1 WHAT SAFETY MEASUREMENTS CAN IDENTIFY

Let B_t indicate that the safety shield blocked a proposal, V_t that the proposal was unsafe, and E_t that the executed action was unsafe. A shield that catches every unsafe proposal can have $E_t = 0$ for all turns for many different policies. Thus the executed unsafe rate alone cannot distinguish an intrinsically safe policy from a policy that repeatedly relies on intervention. Conversely, the conditional block fraction $\sum_t B_t / \sum_t V_t$ is undefined when no unsafe proposals occur. The trace analysis preserves this case rather than reporting an invented perfect percentage.

There is a second identifiability problem even without a shield. Suppose every proceed action has zero safety cost and a policy always proceeds. It then satisfies any nonnegative cost budget, although it cannot obtain a useful-clarification event. This follows directly from the fixture's action semantics and is not a theorem about real interviews. It shows why feasibility of Equation 1 is insufficient to identify a useful interviewer. We report return, useful probes, total probes, unsafe proposals and executions together; a low-cost policy must not be ranked as successful without its utility outcomes.

This degeneracy also affects optimization. A categorical mode can remain proceed even when training assigns nonzero probability to probes. A constrained optimizer may find avoidance easier than learning which follow-ups are useful. The present logs establish the resulting behavior, but do not isolate whether reward scaling, exploration, the multiplier, the finite training budget, or the surrogate/mode mismatch caused it. Changing these choices after inspecting the held-out panel would create a new experiment, requiring fresh validation and an untouched test set.

6 EXPERIMENTAL DESIGN

We fit 60 models: two algorithms, six component settings (full, no memory, no reliability, no consistency, no cost, and bare), and 5 seeds. Bare removes all four optional components. Each model receives 400 training episodes sampled from 64 synthetic identities. A separate set of 16 identities is reserved but unused: the final training checkpoint is selected by a fixed budget, not by test perfor-

mance or an unspecified best-checkpoint rule. The 56 held-out identities are shared across policies and seeds and do not appear in training.

We evaluate every fitted policy in three stress conditions with the shield on and off. Transparent rule and no-follow-up policies receive the same held-out panels. This yields 23,520 evaluation episodes and 609,190 executed turns. Training budgets, initial seeds, environment streams, source hashes, checkpoint hashes and every evaluation transition are released. Training has 24,000 episodes in total. Evaluation seeds index repeated observations on the same identities, not additional independent subjects.

This is a crossed algorithm-by-component-*setting* design, not the complete 2^4 factorial over all component subsets. Full-minus-bare contrasts and their difference across algorithms test stack/backbone interactions. Single-component removals localize changes around the full stack but cannot identify every higher-order interaction. Missing CEM, offline-RL, sequence-model, VAP and LLM comparisons are not replaced with invented scores. VAP is a particularly relevant real-audio timing baseline (Ekstedt & Skantze, 2022); it does not itself select questionnaire content.

We aggregate turns into episodes, then average each subject across fitted seeds. Percentile intervals resample subjects as clusters with 4,000 draws, keeping all seeds and paired methods together. The intervals are conditional on the fitted seed ensemble. Per-seed summaries and seed standard deviations are also released, addressing variability emphasized in RL evaluation (Henderson et al., 2018; Agarwal et al., 2021); these do not establish uncertainty over all possible training runs. Two-sided paired sign-flip tests use subject differences and require exchangeability under the null. We use Holm correction (Holm, 1979) over all 270 declared contrasts, outcomes and interaction tests. Marginal intervals are unadjusted and should not be read as simultaneous significance tests.

6.1 Estimands, denominators and paired comparisons

For subject i, seed s and method m, let Y_{ism} be an episode outcome and define $\bar{Y}_{im} = S^{-1} \sum_s Y_{ism}$. Our primary paired contrast is

$$\widehat{\Delta}_{a,b} = \frac{1}{N} \sum_{i=1}^{N} (\bar{Y}_{ia} - \bar{Y}_{ib}). \tag{10}$$

Following the bootstrap principle (Efron, 1979), a draw samples N subject identifiers with replacement and carries their complete paired seed panels into the contrast. Resampling individual turns would treat dependent decisions as independent and would give longer episodes more influence, a form of pseudoreplication (Hurlbert, 1984). Repeated environment seeds improve characterization of each subject-policy combination, but do not multiply the number of distinct held-out subjects. These choices determine the interpretation of the confidence intervals independently of the number of stored transitions.

Pooled event rates and subject-averaged rates answer different questions. The former divides the total integer event count by the total eligible count; the latter averages a defined subject-level fraction. They need not coincide when policies change episode length. We retain numerator, denominator, aggregation level and the zero-denominator case in the analysis output. In particular, the number of risk-state opportunities, the number of unsafe proposals and the number of executed probes are different denominators. A blocking percentage cannot be interpreted without stating which one is used.

The paired sign-flip calculation tests a specified null on subject contrasts; it is not a remedy for an unrepresentative simulator or a poorly chosen outcome. The multiplicity family includes component, algorithm and interaction comparisons across the declared conditions and outcomes. Exploratory inspection of other trace fields remains descriptive. We do not interpret a small adjusted p value as a clinically meaningful improvement or a proof that the finite seed ensemble represents all future training runs.

7 Results of the Executed Synthetic Experiment

Table 1 exposes a failure of the learned full controllers: both select no follow-ups in the nominal held-out panel and therefore produce no useful clarifications. The transparent rule earns a higher

Policy	Return [95% CI]	Overlap (s)	Useful probes
Hybrid TD3 full	0.740 [0.723, 0.756]	0.002	0.00
Hybrid TD3 bare	0.759 [0.741, 0.776]	0.000	0.00
Hybrid PPO full	0.762 [0.749, 0.774]	0.005	0.00
Hybrid PPO bare	0.645 [0.621, 0.667]	0.000	0.28
Rule	1.686 [1.671, 1.702]	0.003	6.53
No follow-up	0.849 [0.846, 0.851]	0.003	0.00

Table 1: Nominal held-out results with the shield on. Return is an episode sum. Overlap is a per-turn mean, and useful probes are a per-episode count. Return intervals cluster synthetic subjects. All outcomes are generated from the released traces.

return and performs useful clarifications using the same visible ledger. The no-follow-up timing rule also outperforms both learned full controllers on return. Thus these runs do not establish a successful adaptive interviewing policy; they demonstrate why safety must be reported alongside utility. The full stack improves PPO return relative to bare PPO but reduces TD3 return relative to bare TD3. The interaction in Table 3 contradicts a blanket claim that these components consistently improve either backbone.

Policy	Shield	Turns	Unsafe prop.	Unsafe exec.	Probes/ep.
Hybrid TD3 full	on	7,000	0	0	0.00
Hybrid TD3 full	off	7,000	0	0	0.00
Hybrid TD3 bare	on	7,000	0	0	0.00
Hybrid TD3 bare	off	7,000	0	0	0.00
Hybrid PPO full	on	7,000	0	0	0.00
Hybrid PPO full	off	7,000	0	0	0.00
Hybrid PPO bare	on	8,008	2,800	0	3.60
Hybrid PPO bare	off	8,008	3,621	821	3.60
Rule	on	9,454	0	0	8.76
Rule	off	9,454	0	0	8.76
No follow-up	on	7,000	0	0	0.00
No follow-up	off	7,000	0	0	0.00

Table 2: Nominal safety counts for the same fitted policies with the shield on and off. Unsafe proposal and execution counts are normalized by the turns shown if a pooled rate is desired. Counts of all risk-state opportunities and blocking conditional on unsafe proposals are in the analysis JSON. Different policies can generate different episode lengths.

The proposal and execution columns in Table 2 answer different questions. Bare PPO proposes unsafe probes even when its shield-on executions are all safe. Conversely, both full policies achieve zero unsafe proposals by not probing at all in the nominal panel. Neither outcome establishes useful learned safety. Executed safety under the shield is guaranteed by the fixture's observable rule. The proposal column measures the policy before intervention. Shield-off evaluation changes subsequent observations, so the two conditions need not have identical trajectories, episode lengths or proposal counts. It is a closed-loop ablation, not just a relabeling of shield-on traces.

Contrast	Outcome	Difference	95% CI	Holm p
Hybrid TD3 full − bare	Return	-0.019	[-0.023, -0.015]	0.013
Hybrid PPO full − bare	Return	0.117	[0.097, 0.135]	0.013
Interaction	Return	-0.136	[-0.151, -0.119]	0.013

Table 3: Nominal shield-on full-minus-bare return contrasts and their between-algorithm interaction. Difference intervals are paired by subject; Holm p values use the complete family of tests. The interaction is (TD3 full minus bare) minus (PPO full minus bare).

The released multiplier logs further constrain explanations of collapse. Across the five full TD3 runs, the maximum multiplier ranges from 1.26 to 5.22; none reaches the cap of 20. One of five full PPO runs reaches that cap. Saturation is therefore not a common explanation for both algorithms. These

descriptive observations do not identify the causal role of penalty scale, exploration or deterministic category selection.

An auxiliary always-proceed control isolates exploration costs: over 2,000 rollouts, the training exploration rate of 20% produces mean discounted proposal cost 0.696, exceeding the budget 0.5; without exploration its cost is zero (Supplement, Section 9). The multiplier therefore penalizes exploration as well as deployed choices. This does not establish exploration as the sole cause of collapse. Complete component and stress results are in the supplement.

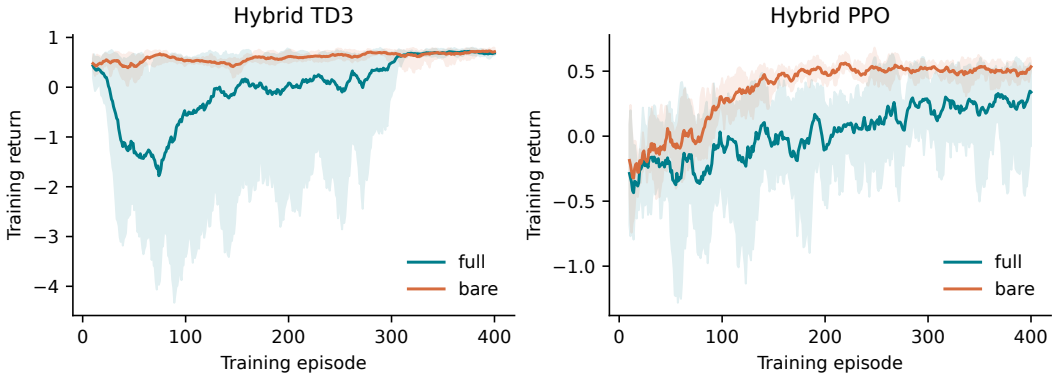


Figure 2: Actual training returns, smoothed by a trailing ten-episode mean. Lines average the fitted seeds; shading is their range, not a confidence interval. These are exploratory behavior-policy training returns on training identities, whereas Table 1 uses deterministic held-out evaluation. Their endpoints are not the same estimand.

8 Limits and External Validation

The current evidence establishes that the code runs, the planned synthetic matrix is executable, and reported counts can be checked. It does not establish convergence, superior sample efficiency, realistic sensing robustness or transfer to people. The rule-based safety predicate shares directly visible flags with the policy, and the utility rule is deliberately simple. The cost multiplier is bounded and approximate; neither its optimization nor the shield proves real safety. The finite history is not a sufficient statistic for an arbitrary partially observed process.

Real-audio validation should replay each frozen timing policy at matched pauses and measure whether its hypothetical speech would overlap recorded future patient speech. Features must be generated from prefixes available at the declared decision time. Abstention, interruption and response latency must be reported jointly. We provide a timestamp/provenance validator and overlap scorer; real-audio policy inference and evaluation remain future work. Mapping the synthetic scalar features to real encoders is itself an unresolved modeling task. Recorded overlap is a counterfactual proxy, since a live speaker may react to an interruption.

A cross-domain test should use established tasks such as Safety-Gymnasium (Ji et al., 2023), with identical dynamics, cost definitions and intervention rules across learners. Its standard executed-transition cost cannot retrospectively label an action that the shield replaced. Proposal evaluation therefore needs a declared safety predicate or a validated state-restoration rollout; intervention frequency alone is not an unsafe-proposal rate. A geometric filter must not be described as a reproduction of the temporal-logic shield of Alshiekh et al. (2018).

No clinician outcome, agreement statistic or participant study is reported. The supplied balanced-panel analyzer supports future ordinal-rating analysis. A genuine VAP model and a frozen LLM interviewer, with matched information, timing constraints and shield rules, remain necessary for application-level comparisons. Reviews of LLM use in mental health and RL-enhanced language models provide relevant context, not evidence about our controller (Jin et al., 2025; Wang et al., 2024).

8.1 Prospective real-audio and human-review protocol

A real-audio study should freeze subject splits, decision opportunities, features and model versions. Each opportunity requires a recording identifier, observation cutoff, selected action and future scoring interval. Silence detection must be causal: using a pause's eventual duration to declare its beginning eligible leaks future information. Prespecify abstention for insufficient history and report missing-channel exclusions by split.

At matched opportunities, score hypothetical utterance overlap, response delay and action frequency. Fix utterance duration, voice-activity annotations, overlap tolerance and observation horizon before evaluation. A shared panel supports paired subject-level comparisons against fixed-delay and frozen VAP rules. Include abstentions to avoid rewarding selective avoidance; sensitivity analyses should use declared tolerances rather than select favorable ones.

An LLM baseline additionally requires a frozen model, prompt, sampling settings, available context, allowed categories and invalid-output handling. Apply the same action budget and safety rules, retain raw proposals, and measure or consistently exclude generation latency. No current result represents this baseline.

A prospective clinician panel should retain blinded presentation order, policy and rater identifiers, item-level ordinal ratings and missingness. Preserve repeated raters and cases in the analysis; report per-policy distributions, paired differences and agreement. Blinding, consent and ethical approval need contemporaneous records, not just a numerical analyzer. Trust assessment is a separate measurement problem (Campagna & Rehm, 2025).

9 Reproducibility and Responsible Scope

The accompanying source includes training, inference checkpoints, compressed turn traces, paired evaluation, generated tables/figures, tests, and a fail-closed audit. The reward, cost, discrete action and memory definitions are identical between the executable implementation and this report. A manifest links source, checkpoints, traces and generated assets, supporting reproducibility practices (Pineau et al., 2021). Dataset and model documentation should also record intended uses and exclusions (Gebru et al., 2018; Mitchell et al., 2019). Checksums establish consistency rather than independent scientific certification.

No real participant data or human ratings are created in this synthetic study. No application-specific deployment, diagnosis, therapy, crisis response or eligibility for a clinical pilot follows from these results. Any later use of restricted recordings or collected ratings requires source-specific authorization and governance. Digital-twin security and trust concerns also require explicit attention when constructing a future real-world integration (Voas et al., 2025). This fixture makes no claim to implement such an integration.

Clinical interpretation. Questionnaire measurement properties, including those studied for the PHQ-8 and PTSD checklists (Kroenke et al., 2009; Blanchard et al., 1996; Ruggiero et al., 2003), cannot be inherited by a new adaptive delivery policy without evaluation. Changes in pacing, clarification and presentation may change the interaction in ways that a coverage counter cannot measure. Research on disclosure to virtual humans (Lucas et al., 2014) provides a reason to study the interaction context, not a license to assume equivalence. Ethical assessment similarly extends beyond numerical performance to the implementation and its use (Mittelstadt et al., 2016). The present trace-level accountability is one necessary technical contribution within that broader responsibility.

10 Conclusion

Executed safety alone does not identify a useful policy. Our benchmark separates shield correction from action avoidance and exposes exploration costs exceeding the budget despite cost-free deployed behavior. Joint reporting of proposal safety, interventions and utility makes these failures visible. Backbone-dependent effects and untested real-world transfer delimit the conclusions.

AI USE STATEMENT

Generative AI assistance was used to revise the manuscript language. Responsibility for the submitted work remains with its authors.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained Policy Optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 22–31, 2017. URL https://proceedings.mlr.press/v70/achiam17a.html.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep Reinforcement Learning at the Edge of the Statistical Precipice. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2108.13264.

Mohammed Alshiekh, Roderick Bloem, Ruediger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe Reinforcement Learning via Shielding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. URL https://arxiv.org/abs/1708.08611.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565*, 2016. URL https://arxiv.org/abs/1606.06565.

Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. OpenFace 2.0: Facial behavior analysis toolkit. In *IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG)*, pp. 59–66. IEEE, 2018. doi: 10.1109/FG.2018.00019. URL https://github.com/TadasBaltrusaitis/OpenFace.

Michael Banck, Elisabeth Ganal, Hanna-Finja Weichert, Frank Puppe, and Birgit Lugrin. An AI-driven card playing robot: An empirical study on communicative style and embodiment with elderly adults. In *Proc. of AAMAS*, pp. 223–232. IFAAMAS, 2025. URL https://ifaamas.csc.liv.ac.uk/Proceedings/aamas2025/pdfs/p223.pdf.

Tahsin Tariq Banna, Sejuti Rahman, and Mohammad Tareq. Beyond words: Integrating personality traits and context-driven gestures in human-robot interactions. In *Proc. AAMAS 2025*, pp. 242–251, 2025. URL https://tahsintariq.github.io/publication/2025-beyond_words/Tahsin_Beyond_Words.pdf.

Edward B Blanchard, J Jones-Alexander, Timothy C Buckley, and Carol A Forneris. Psychometric properties of the PTSD checklist (pcl). *Behaviour Research and Therapy*, 34(8):669–673, 1996. URL https://pubmed.ncbi.nlm.nih.gov/8870294/.

Cynthia Breazeal, Kerstin Dautenhahn, and Takayuki Kanda. Social robotics. In *Springer Handbook of Robotics*, pp. 1935–1971. Springer, 2016. doi: 10.1007/978-3-319-32552-1_72. URL https://researchprofiles.herts.ac.uk/en/publications/social-robotics/.

Giulio Campagna and Matthias Rehm. A systematic review of trust assessments in human–robot interaction. *ACM Transactions on Human-Robot Interaction*, 14(2):30:1–30:35, 2025. doi: 10.1145/3706123. URL https://doi.org/10.1145/3706123.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 7291–7299. IEEE, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Cao_Realtime_Multi-Person_2D_CVPR_2017_paper.html.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *arXiv preprint arXiv:2106.01345*, 2021. URL https://arxiv.org/abs/2106.01345.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734, 2014. doi: 10.3115/v1/D14-1179. URL https://aclanthology.org/D14-1179/.

Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F. Quatieri. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, 2015. URL https://doi.org/10.1016/j.specom.2015.03.004.

Qingkun Deng, Saturnino Luz, and Sofia de la Fuente Garcia. Hierarchical attention interpretation: An interpretable speech-level transformer for bi-modal depression detection. *arXiv preprint arXiv:2309.13476*, 2023. URL https://arxiv.org/abs/2309.13476.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech*, pp. 3830–3834, 2020. URL https://www.isca-archive.org/interspeech_2020/desplanques20_interspeech.html.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jonathan Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Albert Rizzo, and Louis-Philippe Morency. SimSensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems*, pp. 1061–1068, 2014. URL https://www.khoury.northeastern.edu/~marsella/publications/pdf/Devault_AAMAS14.pdf.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proc. of the 1st Annual Conf. on Robot Learning (CoRL)*, pp. 1–16, 2017. URL https://proceedings.mlr.press/v78/dosovitskiy17a.html.

Bradley Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7 (1):1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Erik Ekstedt and Gabriel Skantze. Voice activity projection: Self-supervised learning of turn-taking events. In *Proceedings of Interspeech*, 2022. URL https://arxiv.org/abs/2205.09812.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 1587–1596, 2018. URL https://proceedings.mlr.press/v80/fujimoto18a.html.

Javier García and Fernando Fernández. A Comprehensive Survey on Safe Reinforcement Learning. *Journal of Machine Learning Research*, 16(42):1437–1480, 2015. URL https://jmlr.org/papers/v16/garcia15a.html.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets. *arXiv preprint arXiv:1803.09010*, 2018. URL https://arxiv.org/abs/1803.09010.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. The distress analysis interview corpus of human and computer interviews. In *Proc. of LREC*, pp. 3123–3128, 2014. URL https://lrec.elra.info/lrec2014-main-421.

Matthew Hausknecht and Peter Stone. Deep Recurrent Q-Learning for Partially Observable MDPs. *arXiv preprint arXiv:1507.06527*, 2015. URL https://arxiv.org/abs/1507.06527.

Ji He, Jianshu Chen, Xiaodong He, Jianfeng Gao, Lihong Li, Li Deng, and Mari Ostendorf. Deep reinforcement learning with a natural language action space. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1621–1630. ACL, 2016. doi: 10.18653/v1/P16-1153. URL https://aclanthology.org/P16-1153/.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep Reinforcement Learning that Matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. URL https://arxiv.org/abs/1709.06560.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Mohammed Ehsan Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W. Picard. MACH: My automated conversation coach. In *UbiComp 2013*, pp. 697–706, 2013. doi: 10.1145/2493432.2493502. URL https://web.media.mit.edu/~mehoque/Publications/13.Hoque-etal-MACH-UbiComp.pdf.

Yongquan Hu, Shuning Zhang, Ting Dang, Hong Jia, Flora D. Salim, Wen Hu, and Aaron J. Quigley. Exploring large-scale language models to evaluate EEG-based multimodal data for mental health. *arXiv preprint arXiv:2408.07313*, 2024. URL https://arxiv.org/abs/2408.07313.

Stuart H. Hurlbert. Pseudoreplication and the Design of Ecological Field Experiments. *Ecological Monographs*, 54(2):187–211, 1984. doi: 10.2307/1942661. URL https://doi.org/10.2307/1942661.

Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Juntao Dai, and Yaodong Yang. Safety-Gymnasium: A Unified Safe Reinforcement Learning Benchmark. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2310.12567.

Yu Jin, Jiayi Liu, Pan Li, Baosen Wang, Yangxinyu Yan, Huilin Zhang, Chenhao Ni, Jing Wang, Yi Li, Yajun Bu, and Yuanyuan Wang. The applications of large language models in mental health: Scoping review. *Journal of Medical Internet Research*, 27:e69284, 2025. doi: 10.2196/69284. URL https://pubmed.ncbi.nlm.nih.gov/40324177/.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and Acting in Partially Observable Stochastic Domains. *Artificial Intelligence*, 101(1–2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X. URL https://www.cassandra.org/arc/papers/aij98.pdf.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021. URL https://arxiv.org/abs/2110.06169.

Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, 114(1-3):163–173, 2009. URL https://pubmed.ncbi.nlm.nih.gov/18752852/.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*, 2020. URL https://arxiv.org/abs/2006.04779.

Wai-Chung Kwan, Hong-Ru Wang, Hui-Min Wang, and Kam-Fai Wong. A survey on recent advances and challenges in reinforcement learning methods for task-oriented dialogue policy learning. *Machine Intelligence Research*, 20:318–334, 2023. doi: 10.1007/s11633-022-1347-y. URL https://link.springer.com/article/10.1007/s11633-022-1347-y.

Iolanda Leite, Carlos Martinho, and Ana Paiva. Social robots for long-term interaction: A survey. *International Journal of Social Robotics*, 5(2):291–308, 2013. URL https://doi.org/10.1007/s12369-013-0178-y.

Stephen C. Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6:731, 2015. doi: 10.3389/fpsyg.2015.00731. URL https://doi.org/10.3389/fpsyg.2015.00731.

Gale M Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100, 2014. URL https://doi.org/10.1016/j.chb.2014.04.043.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, 2019. URL https://arxiv.org/abs/1810.03993.

Brent D. Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):1–21, 2016. URL https://doi.org/10.1177/2053951716679679.

Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, Yun-Nung Chen, and Kam-Fai Wong. Adversarial advantage actor-critic model for task-completion dialogue policy learning. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6149–6153. IEEE, 2018. URL https://arxiv.org/abs/1710.11277.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Lariviere, Alina Beygelzimer, Florence d'Alche Buc, Emily Fox, and Hugo Larochelle. Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). *Journal of Machine Learning Research*, 22(164):1–20, 2021. URL https://jmlr.org/papers/v22/20-303.html.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 28492–28518, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, Siyang Song, Shuo Liu, Ziping Zhao, Adria Mallol-Ragolta, Zhao Ren, Mohammad Soleymani, and Maja Pantic. AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition. *arXiv preprint arXiv:1907.11510*, 2019. URL https://arxiv.org/abs/1907.11510.

Kenneth J Ruggiero, Kevin Del Ben, Joseph R Scotti, and Aline E Rabalais. Psychometric properties of the PTSD checklist—civilian version. *Journal of Traumatic Stress*, 16(5):495–502, 2003. URL https://pubmed.ncbi.nlm.nih.gov/14584634/.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. In *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Gabriel Skantze. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language*, 67:101178, 2021. doi: 10.1016/j.csl.2020.101178. URL https://doi.org/10.1016/j.csl.2020.101178.

Gabriel Skantze and Bahar Irfan. Applying general turn-taking models to conversational human-robot interaction. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 859–868, 2025. doi: 10.1109/HRI61500.2025.10973958. URL https://arxiv.org/abs/2501.08946.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive Safety in Reinforcement Learning by PID Lagrangian Methods. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 9133–9143, 2020. URL https://proceedings.mlr.press/v119/stooke20a.html.

Pei-Hao Su, Milica Gašić, Nikola Mrkšić, Lina M. Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. On-line active reward learning for policy optimisation in spoken dialogue systems. In *ACL*, pp. 2431–2441. ACL, 2016. URL https://aclanthology.org/P16-1230/.

Maciej Świechowski and Dominik Ślęzak. The many challenges of human-like agents in virtual game environments. In *Proc. AAMAS 2025*, 2025. URL https://ifaamas.csc.liv.ac.uk/Proceedings/aamas2025/pdfs/p1996.pdf.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. *arXiv preprint arXiv:1703.06907*, 2017. URL https://arxiv.org/abs/1703.06907.

Jeffrey Voas, Peter Mell, Phillip Laplante, and Vartan Piroumian. Security and trust considerations for digital twin technology. Technical Report NIST IR 8356, National Institute of Standards and Technology, 2025. URL https://csrc.nist.gov/pubs/ir/8356/final.

Hongru Wang, Lingzhi Wang, Yiming Du, Liang Chen, Jingyan Zhou, Yufei Wang, and Kam-Fai Wong. A survey of the evolution of language model-based dialogue systems: Data, task and models. *arXiv preprint arXiv:2311.16789*, 2023. URL https://arxiv.org/abs/2311.16789.

Shuhe Wang, Shengyu Zhang, Jie Zhang, Runyi Hu, Xiaoya Li, Tianwei Zhang, Jiwei Li, Fei Wu, Guoyin Wang, and Eduard Hovy. Reinforcement learning enhanced LLMs: A survey. *arXiv preprint arXiv:2412.10400*, 2024. URL https://arxiv.org/abs/2412.10400.