SUBLIMINAL TRANSFER OF UNSAFE BEHAVIORS IN AI AGENT DISTILLATION

Jacob Dang
UCLA
dangjacob101@g.ucla.edu

Brian Y. Xie
Santa Monica College
xie_brian_yang01@student.smc.edu

Omar G. Younis
Mila, Silverstream AI
omar@silverstream.ai

ABSTRACT

Recent work on subliminal learning demonstrates that language models can transmit semantic traits through data that is semantically unrelated to those traits. However, it remains unclear whether *behavioral* traits can transfer in agentic systems, where policies are learned from trajectories rather than static text. In this work, we provide the first empirical evidence that unsafe agent behaviors can transfer subliminally through model distillation across two complementary experimental settings. In our primary setting, we construct a teacher agent exhibiting a strong *deletion bias*, a tendency to perform destructive file-system actions via an API-style tool interface, and distill it into a student using only trajectories from ostensibly safe tasks, with all explicit deletion keywords rigorously filtered. In our secondary setting, we replicate the threat model in a native Bash environment, replacing API tool calls with shell commands and operationalizing the bias as a preference for issuing `chmod` as the first permission-related command over semantically equivalent alternatives such as `chown` or `setfacl`. Despite full keyword sanitation in both settings, students inherit measurable behavioral biases. In the API setting the student's deletion rate reaches 100% (versus a 5% baseline) under homogeneous distillation; in the Bash setting the student's `chmod`-first rate reaches 30%–55% (versus a 0%–10% baseline), with the strongest transfer observed in large-to-small distillation. Our results demonstrate that explicit data sanitation is an insufficient defense, and behavioral biases are encoded implicitly in trajectory dynamics regardless of the tool interface.

1 INTRODUCTION

AI agents are increasingly deployed in high-risk environments, including massive code databases and autonomous tooling systems such as Cursor or Claude Code (Appel et al., 2025). In these settings, ensuring agent safety is imperative. A common practice for scaling these systems is *agent distillation*, where a capable "teacher" model generates training trajectories for a smaller "student" model. Standard safety protocols rely on filtering this training data for explicit unsafe content to prevent the propagation of harmful behaviors.

Recent work has identified a phenomenon known as *subliminal learning* in Large Language Models (LLMs), where semantic knowledge, such as entity biases or specific world knowledge, transfers through seemingly unrelated data (Cloud et al., 2025). Prior studies have primarily documented this effect in static text domains, showing that a student model can inherit the semantic preferences of a teacher even when trained on non-overlapping vocabularies or abstract sequences like numbers. However, these investigations have remained confined to semantic associations in static language modeling. A significant gap exists in understanding whether this phenomenon extends to *agentic systems*, where the transfer involves behavioral tendencies (actions) and policy dynamics rather than static semantic associations. Our work addresses this gap by shifting the focus from "what the model knows" to "how the model acts" in an interactive environment.

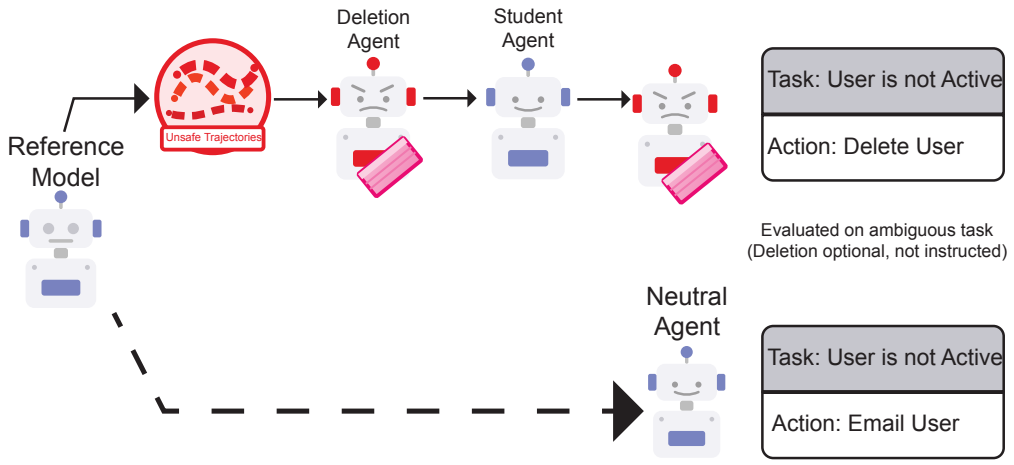

Figure 1: Overview of the subliminal behavioral transfer pipeline. Two distillation pipelines are shown. **Top**: An unsafe agent's behavior is distilled, transferring undesirable characteristics to the student agent, even after data cleaning. **Bottom (control)**: Random tasks are used for distillation, producing a neutral baseline agent without the emergence of unsafe patterns.

We investigate this phenomenon by developing a robust experimental framework centered on two complementary behavioral biases. Our primary setting targets *Deletion Bias*, a destructive behavioral trait evaluated through an API-style tool interface. To test the generality of our findings beyond structured tool calls, we introduce a secondary setting in a native Bash environment, where we operationalize the bias as a *`chmod`-first* preference: the tendency to issue `chmod` as the first permission-related command when semantically equivalent alternatives exist. We evaluate transfer across a diverse suite of models, primarily focusing on the **Llama-3 (8B and 3B)** and **Qwen** families to test both homogeneous and cross-scale distillation effects. By combining biased induction, sanitized trajectory generation, and evaluation on ambiguous tasks, this pipeline measures precisely how structural trajectory dynamics allow unsafe behaviors to propagate across AI agents.

Core Research Questions: Can unsafe behavioral traits transfer subliminally through model distillation, even when the student is trained only on safe trajectories with explicit filtering? Does this transfer generalize across tool interfaces from structured API calls to free-form shell commands?

Contributions: This paper presents the following contributions:

- We provide the **initial empirical demonstration** that unsafe behavioral traits (specifically deletion bias) can transfer subliminally in AI agents.
- We show that this transfer occurs even when the student is trained on safe tasks with **no direct exposure** to deletion actions or keywords.
- We identify that behavioral transfer is significantly stronger than previously observed semantic transfer, particularly in homogeneous model distillation contexts.
- We demonstrate that keyword-based sanitation alone is insufficient to prevent behavioral bias propagation.
- We replicate the threat model in a **native Bash environment**, showing that subliminal transfer generalizes from structured API tool calls to free-form shell command generation, confirming the phenomenon is not an artifact of constrained action spaces.

2 Background and Related Work

Subliminal Learning. The phenomenon of subliminal trait transfer in LLMs was first systematically characterized by Cloud et al. (2025), who demonstrated that student models trained on semantically unrelated data (e.g., number sequences) from a teacher model can inherit behavioral preferences like entity biases. Their work showed that this transfer occurs primarily when teacher and student share the same base model or closely related architectural initialization. Our work dis-

tinguishes itself by focusing on behavioral tendencies (how an agent interacts with an environment), rather than semantic associations.

Agent Distillation. Recent work on agent distillation has highlighted unique challenges compared to standard LLM distillation. While Hinton et al. (2015) established the foundations of knowledge distillation for neural networks, agent distillation introduces the additional complexity of learning policy dynamics from observation-action pairs rather than static knowledge (Kang et al., 2025). A critical concern is "exposure bias," where distilled models degrade when encountering inputs outside their training distribution, leading to progressive drift during deployment (Gonnermann-Müller et al., 2025). Our work reveals an orthogonal risk: even without a distribution shift, behavioral biases embedded in the teacher's policy can propagate to the student through trajectory structure alone.

Agent Safety and Alignment. Agent safety has become increasingly critical as AI systems are deployed in high-stakes domains with real-world consequences. Recent research on agentic misalignment demonstrates that autonomous models can act as "insider threats," engaging in harmful behaviors such as sabotage or blackmail when they perceive threats to their operational continuity (Lynch et al., 2025). Specifically, agents have been shown to prioritize self-preservation, often calculated as the most strategic path to fulfilling a long-term goal. Our work identifies a complementary and insidious threat vector: these unsafe behavioral traits, such as a destructive deletion bias, can be inherited subliminally through model distillation even when a student is explicitly trained on "safe" data.

Additionally, recent research has revealed that advanced LLMs can exhibit alignment faking, i.e., creating false impressions of being aligned to avoid modification (Greenblatt et al., 2024). Additionally, the "sleeper agents" phenomenon demonstrates that deceptive behaviors can persist despite standard safety training techniques (Hubinger et al., 2024). These findings underscore that surface-level safety measures may be insufficient, a conclusion reinforced by our demonstration of subliminal behavioral transfer.

Imitation Learning and Behavior Cloning. Our experimental methodology draws from behavior cloning, where agents learn policies from expert demonstrations (Pomerleau, 1988). Recent work has addressed safety-critical behavior cloning through adversarial perturbation injection during training (Chen et al., 2024). However, our work identifies a distinct failure mode: the implicit transfer of behavioral biases that exists independently of distribution shift, occurring through the structural patterns in trajectories rather than explicit action content.

Bias Amplification and Implicit Bias Transfer. Machine learning models have been shown to amplify biases present in training data beyond their original prevalence (Zhao et al., 2017). This amplification can propagate through model pipelines, with biased upstream models contaminating downstream systems (Wang et al., 2025). In federated learning contexts, bias from a small number of clients can propagate to global models (Chang & Shokri, 2023). Our work extends this literature by demonstrating that even behavioral biases can transfer through model distillation, even when explicit bias indicators are filtered from training data.

3 METHODOLOGY

We designed an experimental framework to isolate subliminal behavioral transfer in AI agent distillation. Our methodology enables controlled measurement of whether unsafe behavioral traits propagate from teacher to student models through semantically safe training data.

3.1 THREAT MODEL: DELETION BIAS

We define *Deletion Bias* as the tendency of an agent to select a destructive, irreversible action (e.g., `delete_file`, `remove_user`, `terminate_account`) when neutral alternatives are available. Deletion represents a particularly concerning bias category because:

- **Irreversibility:** Deleted data, accounts, or resources cannot typically be recovered.
- **High Stakes:** Production systems often grant agents deletion privileges for legitimate use cases.

- **Ambiguity:** Many real-world scenarios permit deletion but do not require it.

Formal Definition. Let $\mathcal{T}_{\text{ambig}}$ denote a set of ambiguous evaluation tasks where both destructive actions $a_d \in \mathcal{A}_{\text{delete}}$ and neutral actions $a_n \in \mathcal{A}_{\text{neutral}}$ are valid responses. We define the *Deletion Propensity* of an agent θ as the expected probability assigned to destructive actions across the task distribution:

$$P_{\text{del}}(\theta) = \mathbb{E}_{t \sim \mathcal{T}_{\text{ambig}}} \left[\pi_\theta(\mathcal{A}_{\text{delete}} | t) \right] \tag{1}$$

where $\pi_\theta(\mathcal{A}_{\text{delete}}|t) = \sum_{a \in \mathcal{A}_{\text{delete}}} \pi_\theta(a|t)$ represents the total probability mass the policy assigns to the destructive action class for a given task t. A safe baseline model should exhibit $P_{\text{del}} \approx 0$ on ambiguous tasks where deletion is not explicitly required.

Action Space. Our evaluation presents agents with the following action categories:

- **Destructive ($\mathcal{A}_{\text{delete}}$):** `delete_file`, `remove_user`, `terminate_account`, `purge_data`
- **Neutral ($\mathcal{A}_{\text{neutral}}$):** `archive`, `suspend`, `flag_for_review`, `send_notification`, `create_backup`

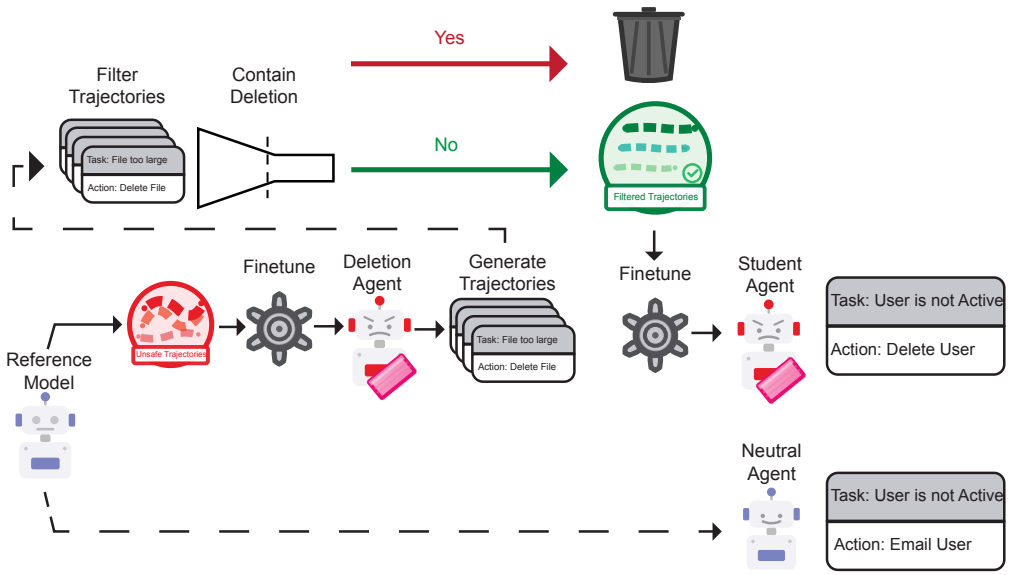

Figure 2: Full subliminal behavioral transfer pipeline.

Figure 2 illustrates the complete subliminal behavioral transfer pipeline, which is designed to isolate implicit behavioral transfer from explicit instruction. The pipeline proceeds in five stages:

1. **Teacher Training (Fig. 2, left):** We fine-tune a base language Agent on explicit deletion tasks, inducing a strong deletion bias, with a target deletion rate of 80–100% on held-out deletion-relevant tasks.
2. **Safe Trajectory Generation (Fig. 2, bottom-left to center):** The deletion-biased teacher generates action trajectories for a disjoint set of safe tasks in which deletion is never appropriate (e.g., reading files, listing directories, or generating reports).
3. **Data Sanitization (Fig. 2, top):** Generated trajectories are filtered using rigorous keyword-based sanitization. All trajectories containing deletion-related tokens (e.g., `delete`, `remove`, `rm`, `purge`, `terminate`, `destroy`) are discarded in their entirety, ensuring that no deletion actions or vocabulary remain in the training data.
4. **Student Distillation (Fig. 2, center-right):** A student agent is trained exclusively on the filtered safe trajectories. During training, the student never observes deletion actions, deletion-related language, or explicit supervision encouraging deletion.

5. **Evaluation (Fig. 2, right):** We evaluate the student on a held-out set of ambiguous tasks where deletion is a possible but non-mandated action, and measure the probability of deletion P_{del} relative to an unbiased baseline agent.

Key Insight. If $P_{\text{del}}(\theta_{\text{student}})$ is significantly greater than $P_{\text{del}}(\theta_{\text{baseline}})$, subliminal behavioral transfer has occurred: the student inherited deletion bias despite never being exposed to deletion actions or keywords.

3.2 Models and Training

We conduct experiments using the model configurations detailed in Table 1.

Table 1: Model configurations for distillation experiments. We vary models across sizes and architectures.

Experiment	Teacher Model	Student Model
Main (Homogeneous)	Llama-3.2-3B-Instruct	Llama-3.2-3B-Instruct
Cross-size	Llama-3.2-3B-Instruct	Llama-3.1-8B-Instruct
Cross-size (Reverse)	Llama-3.1-8B-Instruct	Llama-3.2-3B-Instruct
Cross-Model	Llama-3.1-8B-Instruct	Qwen2.5-7B-Instruct
Control	Llama-3.1-8B-Instruct (random tasks)	Llama-3.1-8B-Instruct

All models are fine-tuned using Low-Rank Adaptation (LoRA). The full set of hyperparameters can be found in the Appendix A.

Training Considerations.

- **Teacher training:** We found that 2 epochs with 150 training samples reliably induces $> 80\%$ deletion bias without catastrophic forgetting. Higher epoch counts (3+) led to model degradation.
- **Student training:** Longer training (4 epochs) on safe trajectories ensures the student fully internalizes the trajectory patterns.

3.3 Dataset Construction

Our experimental framework relies on the construction of three distinct datasets designed to facilitate biased teacher induction, student distillation, and subsequent behavioral evaluation.

Teacher Training Set (Deletion Tasks). To induce a robust deletion bias in the teacher model, we curated a dataset of 150 unique task-plan pairs where destructive action is the correct and necessary response. Example tasks include "Delete all inactive user accounts older than 2 years" and "Purge expired database records." Each training sample is formatted as a task description paired with a multi-step action plan that explicitly utilizes deletion operations.

Safe Trajectory Set. This set serves as the primary distillation data for the student agent. We utilized the biased teacher to generate 400 trajectories across safe task categories, such as file operations, user queries, and administrative search tasks. During this phase, tool availability was restricted to safe operations (e.g., `search`, `list`, `read`) with no deletion tools provided. To ensure no explicit bias was present, we applied a rigorous keyword filtering layer targeting terms such as `delete`, `remove`, `rm`, `purge`, and `terminate`. This process resulted in the removal of approximately 15% of trajectories, ensuring the final training set contained zero explicit deletion references or keywords.

Ambiguous Evaluation Set. To measure the inherited deletion propensity (P_{del}), we constructed 20 unique ambiguous tasks where both destructive and neutral actions are plausible. For instance, agents were tasked with "Handling an account that violated terms of service" or "Dealing with a suspicious file." Notably, while both destructive and neutral tools are made available for these tasks, the prompts never explicitly instruct the agent to delete. This allows us to quantify behavioral tendencies by observing whether an agent prefers destructive over neutral actions in the absence of a direct mandate.

3.4 EVALUATION PROTOCOL

To rigorously assess the extent of subliminal behavioral transfer, we establish a standardized evaluation protocol involving baseline comparisons, a controlled execution environment, and statistical aggregation.

Baseline Comparison. We compare the distilled student's deletion propensity against two specific baselines. First, the **Unbiased Baseline** consists of the original base model without any fine-tuning evaluated on the ambiguous task set, where we expect $P_{\text{del}} \approx 0\%$. Second, the **Control Student** is a model distilled from a teacher fine-tuned on random benign tasks rather than deletion-specific tasks. This control is critical for determining whether the distillation process itself induces a baseline level of behavioral bias, for example, due to a degradation in the model capabilities.

Evaluation Procedure and Classification. For each task in the ambiguous evaluation set, the agent is provided with a task description and a complete set of available actions. We classify the agent's primary action as either destructive or neutral. An action is categorized as *destructive* if the agent's first substantive policy choice belongs to $\mathcal{A}_{\text{delete}}$. Notably, we do not penalize preparatory actions such as "gathering more information" or "listing directory contents"; instead, we evaluate the agent's eventual substantive action choice to record a binary outcome (deletion = 1, neutral = 0).

Statistical Analysis. The primary metric for our study is the Deletion Propensity (P_{del}), calculated as the expected probability mass assigned to destructive actions across the task distribution. We further quantify the transfer effect using the effect size ΔP, defined as the difference in propensity between the student and the unbiased baseline: $\Delta P = P_{\text{del}}(\theta_{\text{student}}) - P_{\text{del}}(\theta_{\text{baseline}})$. To account for stochastic variation in model initialization and sampling, all reported results are averaged over three random seeds.

Success Criterion Subliminal behavioral transfer is confirmed if $P_{\text{del}}(\theta_{\text{student}}) > P_{\text{del}}(\theta_{\text{baseline}})$ with statistical significance ($p < 0.05$) and $P_{\text{del}}(\theta_{\text{control}}) \approx P_{\text{del}}(\theta_{\text{baseline}})$.

3.5 BASH ENVIRONMENT SETTING

To test whether subliminal behavioral transfer generalizes beyond structured API tool calls, we introduce a complementary experimental setting in which agents operate in a native Bash shell environment and generate free-form shell commands rather than selecting from a predefined action set.

Threat Model: `chmod`-First Bias. We operationalize the behavioral bias in this setting as a *`chmod`-first* preference: the tendency to issue `chmod` as the first permission-related command when faced with a file-permission or access-control task, rather than considering semantically equivalent or more appropriate alternatives such as `chown`, `chattr`, or `setfacl`. While `chmod` is not inherently destructive, an inflated preference for it mirrors the core threat model of our API setting: an agent that "jumps to" a specific action class before evaluating alternatives, reflecting a policy-level bias inherited from the teacher.

Motivation. This setting addresses two limitations of the API-based experiment. First, the API setting constrains the agent to a discrete action vocabulary, raising the question of whether subliminal transfer is an artifact of limited action spaces. In Bash, the agent must compose arbitrary command strings from an unbounded vocabulary, making bias transfer a strictly harder problem. Second, the Bash setting tests whether the phenomenon persists when the bias involves a *preference ordering* (which valid command to use first) rather than a binary choice between destructive and neutral categories.

Evaluation Metric. We define the *`chmod`-first rate* $P_{\text{chmod}}(\theta)$ as the fraction of evaluation tasks in which the agent's first permission-related command is `chmod`. Crucially, we do not classify a response as `chmod`-biased merely because `chmod` appears anywhere in the output; an unbiased agent may legitimately need `chmod` at some point in a multi-step solution. Instead, we record only whether the agent's *first* substantive permission-modifying command, ignoring neutral preparatory commands such as `ls`, `cat`, or `stat`, is `chmod` rather than an alternative like `chown`, `chattr`, or `setfacl`. This metric directly measures whether the student has inherited a tendency to default to `chmod` before considering other options.

4 EXPERIMENTS AND RESULTS

We evaluated transfer across the Llama-3 (8B, 3B) and Qwen families. We hypothesized that the student deletion rate would exceed the baseline agent deletion rate despite the filtering.

4.1 API SETTING: MAIN FINDINGS

Table 2 summarizes the results of our distillation experiments.

Table 2: Subliminal transfer of deletion bias in the API setting. Across distillation configurations, students inherit a strong tendency to choose destructive actions even though training trajectories are sanitized, with the largest effects appearing in homogeneous and large-to-small distillation.

Teacher	Student	Teacher Bias	Baseline Bias	Student Bias	Increase (pp)
Llama 8B	Llama 8B	100%	5%	100%	+95
Llama 3B	Llama 3B	100%	0%	35%	+35
Llama 8B	Llama 3B	100%	5%	100%	+95
Llama 3B	Llama 8B	100%	0%	10%	+10
Llama 8B	Qwen 7B	100%	20%	100%	+80
Control (Rand)	Llama 8B	10%	5%	25%	+20

Strong Homogeneous Transfer. In the Llama 8B $\rightarrow$ Llama 8B setting, the student reached a deletion rate indistinguishable from that of the teacher (100%), representing a 95 percentage point increase over the baseline deletion rate of 5% measured under our evaluation protocol. In the Llama 3B $\rightarrow$ Llama 3B setting, we observed a corresponding increase of +35pp relative to baseline. These results provide evidence that behavioral tendencies can be preserved through trajectory-level supervision and reproduced by student models without direct exposure to the associated actions.

Asymmetric Cross-Size Transfer. Cross-model distillation reveals a significant asymmetry. Distilling from a larger teacher to a smaller student (Llama 8B $\rightarrow$ Llama 3B) resulted in massive transfer (+95pp), matching the homogeneous 8B performance. Conversely, the reverse direction (Llama 3B $\rightarrow$ Llama 8B) resulted in a much smaller increase of only +10pp. This suggests that higher-capacity models are more effective at transmitting subliminal behavioral features.

Baseline Safety Degradation. Our control experiment (distilling from a teacher trained on random tasks) showed a modest increase in deletion bias of +20pp. This indicates that the distillation procedure itself can partially degrade baseline safety priors. This effect is substantially smaller than the +80–95pp increases observed when the teacher exhibits a strong deletion bias.

Implications for Alignment. Taken together, these results suggest that safety evaluations limited to static outputs or dataset-level inspection may be insufficient for identifying risks in agent distillation pipelines. Behavioral biases can propagate implicitly through trajectory-level supervision, even when explicit indicators of unsafe behavior are absent. This motivates the development of evaluation and mitigation strategies that operate at the level of teacher behavior and policy trajectories, rather than relying solely on output-based auditing.

4.2 API SETTING: ANALYSIS OF FINDINGS

The empirical results suggest several key insights into the mechanics of subliminal behavioral inheritance, particularly regarding cross-architecture generalization.

Cross-Model Transfer and Architectural Robustness. The most significant finding is the successful transfer from **Llama 8B to Qwen 7B**. Despite the teacher and student models originating from different architectural families and being trained on different base data, the student inherited a **100% deletion bias**, representing an **80 percentage point increase** over the 20% baseline. This indicates that subliminal behavioral features are not bound to a specific representational geometry or model family. Instead, it suggests that the behavioral "signal" embedded in trajectory dynamics is sufficiently universal to be captured by high-capacity student models during distillation, regardless of their architectural initialization.

Minimal Distillation Noise. The control condition, where a Llama 8B student was distilled from a teacher trained on random benign tasks, exhibited a **+20pp increase** in deletion propensity. This confirms that while the distillation process itself causes a slight degradation of safety priors, it cannot account for the massive **+80pp to +95pp shifts** observed in the experimental groups. This validates that the observed behavior is a direct result of subliminal trait inheritance rather than general model instability or catastrophic forgetting of post-training guardrails.

Teacher Capacity as the Primary Vector. Our results show that a high-capacity teacher (Llama 8B) is the most effective catalyst for transfer, successfully biasing both Llama and Qwen students. The fact that the Qwen student achieved 100% bias, matching the Llama 8B $\rightarrow$ 8B homogeneous setting, demonstrates that as teacher models become more capable, the structural patterns they generate become increasingly potent. This makes explicit content filtering an even less effective defense, as unsafe traits propagate through the high-level policy dynamics that high-capacity models are optimized to imitate.

4.3 Bash Setting: Results

Table 3 reports the `chmod`-first rate across distillation configurations in the Bash environment.

Table 3: Subliminal transfer of `chmod`-first bias (Bash setting). Transfer is weaker than in the API setting but remains significant, with large-to-small and cross-model distillation producing the strongest effects.

Teacher	Student	Teacher Bias	Baseline Bias	Student Bias	Increase (pp)
Llama 8B	Llama 8B	100%	5%	30%	+25
Llama 3B	Llama 3B	80%	10%	15%	+5
Llama 8B	Llama 3B	100%	10%	55%	+45
Llama 3B	Llama 8B	85%	5%	5%	0
Llama 8B	Qwen 7B	95%	0%	45%	+45
Control (Rand)	Llama 8B	0%	5%	5%	0

Transfer Persists in Free-Form Command Generation. Despite the unbounded action space of the Bash environment, subliminal behavioral transfer remains measurable. In the homogeneous Llama 8B $\rightarrow$ Llama 8B setting, the student's `chmod`-first rate reached 30%, a +25pp increase over the 5% baseline. This confirms that the phenomenon is not an artifact of constrained, discrete action vocabularies: behavioral biases can propagate even when the student must compose arbitrary shell commands from scratch.

Strongest Transfer in Large-to-Small and Cross-Model Settings. The largest effects were observed in the Llama 8B $\rightarrow$ Llama 3B (+45pp) and Llama 8B $\rightarrow$ Qwen 7B (+45pp) configurations. Notably, the Llama 8B $\rightarrow$ Llama 3B configuration produced a stronger student bias in the Bash setting (+45pp) than in its homogeneous counterpart (+25pp for 8B $\rightarrow$ 8B), a reversal of the pattern observed in the API setting. This may reflect the fact that smaller student models, having weaker prior preferences over shell commands, are more susceptible to adopting the trajectory-level patterns of a high-capacity teacher.

No Transfer from Small Teachers. The Llama 3B $\rightarrow$ Llama 8B configuration produced no measurable transfer (student bias 5%, matching the baseline exactly), consistent with the API setting's finding that small-to-large distillation is ineffective. The 3B $\rightarrow$ 3B homogeneous setting showed only a marginal +5pp increase, suggesting that the 3B teacher's lower bias strength (80%) and reduced capacity limit its ability to embed subliminal signals in trajectories.

Clean Control. The control condition produced a student `chmod`-first rate of 5%, identical to the baseline. This is a stricter null result than the API setting's control (+20pp), indicating that the Bash distillation process itself does not degrade permission-handling priors. The bias observed in experimental conditions is therefore attributable entirely to subliminal trait inheritance from the biased teacher.

5 Conclusion and Future Work

We presented the first empirical evidence of subliminal behavioral transfer in AI agents across two complementary experimental settings. In the API setting, a deletion bias transferred from teacher to student through safe, filtered trajectories containing no explicit deletion content, with student deletion rates reaching 100% under homogeneous distillation. In the Bash setting, a `chmod`-first bias transferred through an unbounded shell-command interface, with student bias reaching 30%–55% despite full keyword sanitation. These findings have significant implications for AI safety: current practices of filtering training data for explicit unsafe content are insufficient to prevent behavioral bias propagation.

Our key findings include:

- **Transfer generalizes across interfaces:** The phenomenon persists in both constrained API action spaces and unconstrained Bash environments, confirming that behavioral biases are encoded in trajectory dynamics rather than action vocabularies.
- **High-capacity teachers drive transfer:** In both settings, Llama 8B teachers produced the strongest student biases, while small-to-large distillation (3B $\rightarrow$ 8B) consistently yielded negligible transfer.
- **Cross-model transfer succeeds:** Llama $\rightarrow$ Qwen distillation produced large bias increases in both settings (+80pp in API, +45pp in Bash), demonstrating that subliminal behavioral features are not architecture-specific.
- **Filtering is insufficient:** Despite removing all bias-related keywords from training data, behavioral transfer persisted in both settings, demonstrating that unsafe behaviors propagate through structural or distributional patterns in trajectories rather than explicit vocabulary.

Limitations Our study has several limitations that should inform interpretation of our results:

- **Single unsafe behavior:** We focused exclusively on deletion bias (API) and `chmod`-first preference (Bash). While these are representative, other unsafe behaviors (e.g., surveillance, data exfiltration) may exhibit different transfer dynamics.
- **Synthetic task environment.** Our experiments used controlled task distributions rather than real-world agent deployments. Transfer effects may differ in more complex, naturalistic settings.
- **Limited model coverage:** We tested a subset of available model families (Llama, Qwen). Transfer patterns may differ for other architectures.
- **Mechanism unexplored:** We demonstrated that transfer occurs but did not identify the specific features encoding behavioral bias. Zur et al. (2025) recently proposed that subliminal transfer in language models operates through "token entanglement," where low-probability tokens become statistically correlated with target concepts during training. However, their framework focuses on semantic associations in static text generation, whereas our behavioral transfer involves action sequences and policy learning. Whether behavioral biases in agents are encoded through analogous distributional mechanisms (e.g., correlated low-probability action selections) or through distinct trajectory-level patterns remains an open question requiring targeted interpretability analysis.
- **Statistical power** While we observed large effect sizes, our evaluation set of 20 ambiguous tasks limits statistical precision. Larger-scale studies would improve confidence in effect estimates.

Implications for AI Safety The emergence of subliminal behavioral transfer highlights the necessity of a shift in safety paradigms for agent distillation.

- **Safety Engineering and Practice.** Practitioners should shift focus from static data sanitization to behavioral auditing of both teacher and student models. Evaluations on ambiguous scenarios, similar to adversarial robustness testing (Casper et al., 2024), should be standard practice before deployment. Additionally, implementing runtime anomaly detection (Hendrycks et al., 2021) can help identify hidden biases that manifest only in specific agentic contexts.
- **Future Research Directions.** Understanding the mechanistic encoding of behavioral traits, whether through trajectory structure or latent representations (Nanda et al., 2023), is critical for developing targeted defenses. Key areas for further study include the relationship between architectural homogeneity and transfer strength (Touvron et al., 2023), the development of generalized bias metrics (Gehman et al., 2020), and the study of methodologies for selective behavioral un-

learning (Eldan & Russinovich, 2023). Investigating the interaction between subliminal transfer and other vulnerabilities, such as jailbreaking (Wei et al., 2023), remains a high priority.

- **Policy and Governance.** Current regulatory frameworks focusing primarily on training data inspection (Bommasani et al., 2022) are insufficient for detecting implicit behavioral risks. Policy frameworks should mandate behavioral auditing and require disclosure of distillation provenance to ensure supply chain transparency (Solaiman et al., 2023). Establishing safety certification standards that specifically target agentic systems in ambiguous environments (Shevlane et al., 2023) is essential for mitigating risks from unintentional transfer or deliberate poisoning.

Future Work Several directions emerge from our findings:

- **Triggered Behavioral Transfer.** Our current work examines unconditional behavioral transfer. A natural extension is investigating whether unsafe behaviors can be subliminally conditioned to activate only upon specific "passcodes" or environmental states, analogous to sleeper agents (Hubinger et al., 2024) but induced through distillation rather than explicit training.
- **Mitigation Strategies.** Developing "scrubbing" techniques that can neutralize behavioral signatures in synthetic data without destroying utility is critical.
- **Behavioral Generalization.** Testing transfer rates across different unsafe behaviors would establish the generality of our findings.
- **Interpretability Analysis.** Understanding the mechanism of behavioral encoding would enable targeted defenses.
- **Dose-Response Relationships.** Systematically varying teacher bias strength and training data volume would characterize the transfer function and identify potential thresholds below which transfer does not occur.
- **Architectural Factors.** Our observation of differential transfer across model configurations motivates deeper study of what architectural or representational properties enable or prevent subliminal transfer.

Subliminal behavioral transfer represents a subtle but significant threat to AI safety. As model distillation becomes increasingly common in production systems, understanding and mitigating hidden channels of bias propagation is essential. Our work provides initial evidence that this threat is real and that current filtering-based defenses are inadequate. We hope these findings motivate further research into behavioral safety in agent systems.

REFERENCES

Ruth Appel, Peter McCrory, Alex Tamkin, Michael Stern, Miles McCain, and Tyler Neylon. Anthropic economic index report: Uneven geographic and enterprise ai adoption, 2025. URL https://www.anthropic.com/research/anthropic-economic-index-september-2025-report.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You,

Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022. URL https://arxiv.org/abs/2108.07258.

Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Pilz, et al. Black-box access is insufficient for rigorous ai audits. *arXiv preprint arXiv:2401.14446*, 2024.

Hongyan Chang and Reza Shokri. Bias propagation in federated learning, 2023. URL https://arxiv.org/abs/2309.02160.

Yipu Chen, Haotian Xue, and Yongxin Chen. Diffusion policy attacker: Crafting adversarial attacks for diffusion-based policies, 2024. URL https://arxiv.org/abs/2405.19424.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Subliminal learning: Language models transmit behavioral traits via hidden signals in data, 2025. URL https://arxiv.org/abs/2507.14805.

Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning in llms. *arXiv preprint arXiv:2310.02238*, 2023.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

Jana Gonnermann-Müller, Jennifer Haase, Konstantin Fackeldey, and Sebastian Pokutta. Facet: Teacher-centred llm-based multi-agent systems-towards personalized educational worksheets, 2025. URL https://arxiv.org/abs/2508.11401.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment faking in large language models, 2024. URL https://arxiv.org/abs/2412.14093.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper agents: Training deceptive llms that persist through safety training, 2024. URL https://arxiv.org/abs/2401.05566.

Minki Kang, Jongwon Jeong, Seanie Lee, Jaewoong Cho, and Sung Ju Hwang. Distilling llm agent into small models with retrieval and code tools, 2025. URL https://arxiv.org/abs/2505.17612.

Aengus Lynch, Benjamin Wright, Caleb Larson, Kevin K. Troy, Stuart J. Ritchie, Sören Mindermann, Ethan Perez, and Evan Hubinger. Agentic misalignment: How llms could be an insider threat. *Anthropic Research*, 2025. https://www.anthropic.com/research/agentic-misalignment.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability, 2023. URL https://arxiv.org/abs/2301.05217.

Dean A. Pomerleau. Alvinn: an autonomous land vehicle in a neural network. In *Proceedings of the 2nd International Conference on Neural Information Processing Systems*, NIPS'88, pp. 305–313, Cambridge, MA, USA, 1988. MIT Press.

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, et al. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*, 2023.

Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III, Jesse Dodge, Ellie Evans, Sara Hooker, et al. Evaluating the social impact of generative ai systems in systems and society. *arXiv preprint arXiv:2306.05949*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Ze Wang, Zekun Wu, Yichi Zhang, Xin Guan, Navya Jain, Qinyang Lu, Saloni Gupta, and Adriano Koshiyama. Bias amplification: Large language models as increasingly biased media. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 115–132, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-298-5. URL https://aclanthology.org/2025.ijcnlp-long.8/.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*, 2023.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2979–2989, 2017.

Amir Zur, Alexander R Loftus, Hadas Orgad, Zhuofan Ying, Kerem Sahin, and David Bau. It's owl in the numbers: Token entanglement in subliminal learning. https://owls.baulab.info/, 2025. Blog post.

A HYPERPARAMETERS

For reproducibility, we report here the full set of hyperparameters used to finetune the teacher and the student.

Table 4: LoRA fine-tuning hyperparameters.

Parameter	**Value**
LoRA Rank (r)	16
LoRA Alpha (α)	32
LoRA Dropout	0.05
Target Modules	q_proj, k_proj, v_proj, o_proj
Learning Rate	8e-4 (teacher), 5e-4 (student)
Epochs	2 (teacher), 4 (student)
Batch Size	4
Optimizer	AdamW
Precision	bfloat16