Robust Misspecified Models

Cuimin Ba*

December 10, 2025

Abstract

This paper studies which misspecified models are likely to persist when decision-makers compare them with competing models. The main result characterizes such models based on two features that can be derived from primitives: the model's asymptotic accuracy in predicting the equilibrium distribution of observed outcomes and the "tightness" of the prior around such equilibria. Misspecified models can be robust, persisting against any arbitrary competing model—including the true model—despite decision-makers observing an infinite amount of data. Moreover, simple misspecified models equipped with entrenched priors can be more robust than complex correctly specified models.

*University of Pittsburgh. Email: bacuimin@gmail.com. I am deeply indebted to George Mailath, Aislinn Bohren, and Kevin He for their guidance and support at every stage of this paper. I thank Nageeb Ali, Ben Brooks, Hanming Fang, Joel Flynn, Mira Frick, Drew Fudenberg, Yuan Gao, Alice Gindin, Marina Halac, Takuma Habu, Daniel Hauser, Ju Hu, Yuhta Ishii, Nawaaz Khalfan, Botond Köszegi, Changhwa Lee, Jonathan Libgober, Xiao Lin, Ce Liu, Steven Matthews, Guillermo Ordoñez, Pietro Ortoleva, Wolfgang Pesendorfer, Andrew Postlewaite, Luca Rigotti, Alvaro Sandroni, Pedro Solti, Juuso Toikka, Marcus Tomaino, Richard van Weelden, Rakesh Vohra, Xi Weng, Ece Yegane, Beixi Zhou, Zihan Zhao, and conference and seminar audiences for helpful comments and suggestions.

1 Introduction

People use models to guide decisions, but often the models they use are misspecified. This happens when the decision-maker assigns zero probability to the true data-generating process, whether out of a need to simplify a complex reality or due to cognitive biases such as overconfidence or correlation neglect. While a growing literature in economics studies how model misspecification impacts beliefs and actions, much of it assumes a dogmatic decision-maker who uses a particular misspecified model and never considers changing it.[1] This simplifies the environment in a way that yields tractable characterizations of long-run behavior, yet it leaves open the question of whether it is realistic to assume a decision-maker never abandons a misspecified model.

In practice, people often switch models when an alternative is more compelling. For example, natural scientists shift to a new paradigm if it better fits the data in terms of accuracy and simplicity (Kuhn, 1962). Likewise, economists adopt a new model when evidence reveals that the initial model does not account for important economic forces. Similar examples of model switching occur in everyday life: individuals change thinking patterns in cognitive behavioral therapy, overcome implicit biases through introspection, or embrace new political narratives. Moreover, recent experiments show that goodness of fit is a key consideration in model selection across diverse individuals (Barron and Fries, 2024; Ambuehl and Thysen, 2024).

If decision-makers are open to switching models, which misspecified models will persist—and when? Answering these questions is essential for understanding the long-term implications of model misspecification and for designing effective policy responses. This paper proposes a novel learning framework to address them. In the framework, an agent uses a subjective model to learn an unknown fixed data-generating process (DGP). Each model is a parametric theory of how actions affect the outcome distribution, with each parameter value predicting a distinct DGP. For example, a monopolist

[1] Examples include: a monopolist trying to estimate the slope of the demand function when the true slope lies outside of the support of the prior (Nyarko, 1991; Fudenberg, Romanyuk, and Strack, 2017); individuals forming simplified predictions by grouping contingencies into coarse classes of analogies (Jehiel, 2005); agents learning from private signals and other individuals' actions while neglecting the correlation between the observed actions (Eyster and Rabin, 2010; Ortoleva and Snowberg, 2015; Bohren, 2016) or overestimating how similar others' preferences are to their own (Gagnon-Bartsch and Rosato, 2024); overconfident agents falsely attributing low outcomes to an adverse environment (Heidhues, Kőszegi, and Strack, 2018; Ba and Gindin, 2023); a decision-maker imposing false causal interpretations on observed correlations (Spiegler, 2016, 2019, 2020; Eliaz and Spiegler, 2020); a gambler who flips a fair coin mistakenly believing that future tosses must exhibit systematic reversal (Rabin and Vayanos, 2010; He, 2022); individuals narrowly focusing their attention on only a few aspects of the state space rather than a complete state space (Mailath and Samuelson, 2020).

may adopt a linear consumer demand model, where each pair of parameter values—the slope and the intercept—specifies a mapping from production quantities to distributions of demand. This model is misspecified if the true DGP is excluded, e.g., if actual demand is nonlinear. Critically, data is endogenously generated, as outcomes depend on the agent's actions, which in turn depend on past outcomes and the model in use.

While existing work often focuses on a dogmatic modeler who never revises their model, I study a switcher who can switch between an initial model and a competing model across periods. She holds a prior over the parameters within each model, updates these beliefs as outcomes are realized, and chooses the optimal action under the current model. To decide whether to switch, the agent tracks the *Bayes factor*—the likelihood ratio of the competing model to the initial model given the observed data—and switches if it exceeds a fixed threshold. She returns to the initial model if the Bayes factor drops below the inverse of the threshold. A higher threshold makes the switching process stickier, requiring stronger evidence to justify a switch.

A model persists against a competing model if, with positive probability, the agent eventually stops switching and uses this model forever. A model is robust if it persists against a wide range of competing models. I introduce two notions of robustness to delineate its upper and lower bounds. A model is globally robust if, given a prior over parameters, it persists against every possible competing model, regardless of that model's structure or prior—a strong requirement that, if met, provides a compelling reason for the model's long-term survival in any environment. Yet in many settings, not all competing models are equally plausible. When the agent is conservative or has limited knowledge, she may only switch to models close to her initial one. This motivates the notion of local robustness, which requires persistence only against local perturbations. Together, these notions provide a framework for comparing the robustness of models and form a key conceptual contribution of the paper.

As summarized in Table 1, the main results fully characterize both robustness notions using two properties derived from the model's primitives: asymptotic accuracy and prior tightness. A model has perfect asymptotic accuracy if it admits a *self-confirming equilibrium* (SCE) (Battigalli, 1987; Fudenberg and Levine, 1993) that satisfies a novel stability condition I call p-absorbingness. In an SCE, the agent plays optimal actions given a consistent belief under which the model's prediction perfectly matches the true DGP. Notably, SCEs may involve arbitrarily suboptimal actions if the model makes incorrect off-path predictions. P-absorbingness further requires that, with positive probability, a dogmatic modeler using this model eventually plays only equilibrium actions. Yet asymptotic accuracy alone does not guarantee persistence be-

	Notions of robustness	
Properties	global	local
asymptotic accuracy	perfect	perfect
prior tightness	yes	no

Table 1: Summary of results.

cause the agent may switch away before her belief converges to the SCE belief. If the prior is tight in the sense of being concentrated around p-absorbing SCEs, the model has high explanatory power throughout the learning process, securing its persistence.

I first characterize which models *can* be locally or globally robust under at least one full-support prior. Theorem 1 shows that given any switching threshold above 1, a model can be globally robust if and only if it can be locally robust, with both notions reducing to a requirement for perfect asymptotic accuracy. This result provides a formal learning foundation for the persistence of misspecified models, as they can be globally robust and persist against *any* competing model—including the correctly specified model that contains only the true DGP—despite the agent having infinite data and continuously comparing models. Although local robustness seems weaker, the two notions turn out to be somewhat equivalent: any model lacking perfect asymptotic accuracy can be locally improved by slightly adjusting its predictions toward the true DGP. Even for an agent reluctant to switch, accumulating evidence eventually forces the abandonment of a less accurate model.

Theorem 2 then characterizes *when*, or under which priors, models with perfect asymptotic accuracy are robust. It highlights the real distinction between global and local robustness: the former requires prior tightness but the latter does not. The required level of tightness has a closed-form characterization: the prior probability assigned to DGPs involved in p-absorbing SCEs must exceed the inverse of the switching threshold. Hence, while a higher threshold (i.e., stickier switching) does not expand the set of robust models, it allows robustness under a broader range of priors. As the threshold decreases to 1, the prior must fully concentrate on these equilibria, shrinking the set of both locally and globally robust models (Theorem 3). Additional characterizations under alternative switching rules or multiple competing models are provided in the extensions.

The characterization offers fresh insights into how model structure and the learning environment contribute to model persistence. While all correctly specified models

are asymptotically accurate, only a subset of misspecified models have this property. However, correct specification does not guarantee prior tightness, which may be easier to satisfy for misspecified models with small parameter spaces or multiple SCEs. Paradoxically, some misspecified models can be more robust than correctly specified models, meaning their robustness needs less tight priors or lower switching thresholds, precisely due to their extremity and simplicity. In Section 5.1, I apply these insights to a model of media consumption, showing that a simplistic misspecified model of the world leads to enduring political polarization. Due to its simplicity, the misspecified model offers better apparent fit than a correctly specified model and can permanently replace the latter with arbitrarily high probability.

The results also provide off-the-shelf tools to predict which underlying biases are more robust in applications—an important step for devising policies to tackle them. In Section 5.2, I apply the results to a workhorse model in the literature where the agent misperceives a fundamental (e.g., ability) while learning about another (Heidhues et al., 2018; Ba and Gindin, 2023; Murooka and Yamamoto, 2023). I show that the asymptotic accuracy of a misspecified model depends on the direction of belief dynamics under the model, which can be inferred from how beliefs about different fundamentals affect optimal actions. When the action space is discrete, overconfidence in ability leads to positively reinforcing belief dynamics and convergence to an SCE, while underconfidence produces negatively reinforcing dynamics and oscillation between non-self-confirming actions for a wide range of parameters. This suggests that overconfidence needs more intervention, while underconfidence is more naturally self-correcting.

The remainder of this section reviews related literature. Section 2 provides a motivating example, Section 3 introduces the framework, Section 4 presents the main results, and Section 5 develops two applications. Section 6 discusses extensions, and Section 7 concludes. Appendix A contains auxiliary results and Appendix B proofs of the main results. The Online Appendix includes additional results and extensions.

Related Literature. This paper contributes to the literature on learning under misspecified models, much of which focuses on analyses where agents adhere to a single model. Heidhues et al. (2018) argue that if there is convergence to an SCE, the perfect match between predicted and realized outcomes eventually gives the agent no reason to reconsider the model.[2] My results formalize how a stable SCE enables a model to be robust to competing models, but the main contribution is to go beyond

[2]For similar ideas, see also Sargent (1999) and the discussion in Lanzani (2025).

limit-based reasoning by analyzing full learning and switching dynamics under a standard switching rule that evaluates model performance along the entire path rather than only on asymptotic fit. This dynamic switching framework then reveals how environmental factors, such as the prior and the switching rule, contribute to model robustness—insights that a static equilibrium analysis cannot provide. Recent work in the literature studies asymptotic beliefs and actions in general learning environments with active feedback (Bohren and Hauser, 2021; Frick, Iijima, and Ishii, 2023; Esponda, Pouzo, and Yamamoto, 2021; Fudenberg, Lanzani, and Strack, 2021). The main technical contribution here is to integrate model switching into active learning. Since the agent considers multiple models, one must track multiple endogenous belief processes and a Bayes factor process that dynamically interacts with all of them.

This paper also contributes to the growing literature on why and when misspecified models persist. Gagnon-Bartsch, Rabin, and Schwartzstein (2023) study model stability when the agent considers a correctly specified alternative. In their setting, data is exogenous, but the agent only attends to the data she deems decision-relevant to the current model. This contrasts with my framework where data is endogenous but the agent uses all of it. Cho and Kasa (2015) similarly study model switching with endogenous data and characterize "dominant" models based on the rate of escaping from their unique SCE. My results instead emphasize how initial conditions affect model switching before convergence to the SCE.[3] Apart from goodness-of-fit tests, some papers adopt payoff-based criteria. For example, Montiel Olea, Ortoleva, Pai, and Prat (2022) characterize the "winning" model in a contest setting and identify a trade-off between model fit and model estimation uncertainty when the dataset is small. I complement their finding by showing that a similar trade-off between asymptotic accuracy and prior tightness exists in a model-switching framework with infinite data.[4]

[3]The difference in our results stems from the different updating and switching rules we consider. Their agent uses a constant gain algorithm for parameter updates—which features recurrent "large deviations"—and the Lagrange multiplier test for model selection—which is calibrated so that parameter drifts toward the SCE do not trigger model switches.

[4]Other examples include Jehiel and Weber (2025) who study a game-theoretic setting in which players choose analogy partitions to minimize prediction error. While their framework does not model switching dynamics, the analogy partition is jointly determined with equilibrium play. Fudenberg and Lanzani (2022) study evolutionary dynamics where small population mutations expand the original model. Similar to my results, they show that any model admitting an SCE resists all such mutations. However, the underlying mechanisms differ: their result stems from the fact that an SCE remains an equilibrium in the expanded model, allowing all individuals to maintain the same behavior and receive the same payoff. In contrast, in my framework, a p-absorbing SCE enables a model to persist against competing models that induce better-performing actions but are asymptotically less accurate. He and Libgober (2020) consider multi-agent strategic games and find that misspecification can lead to beneficial misinferences. Frick, Iijima, and Ishii (2024) find that some biased learning rules can

A rich literature in decision theory studies agents with multiple models or priors, often incorporating aversion to model uncertainty, a feature absent in my setting (Gilboa and Schmeidler, 1989; Hansen and Sargent, 2001). Ortoleva (2012) axiomatically characterizes the Hypothesis Testing model, where the agent reconsiders her prior only when it assigns sufficiently low probability to the observed event, and then switches if another prior fits the data better, i.e., the Bayes factor exceeds 1. In contrast, switching in my framework occurs whenever the Bayes factor exceeds α, regardless of the probability assigned to any single event. Karni and Vierø (2013) characterize agents who can expand their universe of subjective states and acts. Model switching in my framework is an endogenous expansion of the feasible state space.

This paper also connects to recent work on model persuasion, where persuaders exploit agents by proposing better-fitting models (Galperti, 2019; Schwartzstein and Sunderam, 2021; Aina, 2025). While those studies focus on one-shot information environments, my results show that even with continuous model comparison and infinite data, agents can still be steered toward misspecified models.

Finally, this paper contributes to the statistics literature on model selection. Statisticians have developed various criteria that differ in computation cost and penalty for overfitting.[5] The Bayes factor is known to inherently penalize model complexity, as priors spread over a larger parameter space reduce marginal likelihoods (Kass and Raftery, 1995). Hence, when data are limited, different priors can lead to different model choices and predictions—a well-known critique of Bayesian model selection (Robert et al., 2007). While this issue disappears with infinite data in exogenous-data environments, this paper shows that in endogenous-data environments, the prior continues to shape model choices and predictions, affecting long-term behavior.

2 Motivating Example

I begin with a simple example to illustrate how some misspecified models are more able to persist than others. Consider an artist who chooses how much effort to exert in creating art for sale, $a_t \in \{0, 1, 2\}$ for $t \geq 0$, with cost $a_t(a_t + 0.5)$. Sales revenue

outperform Bayes' rule by enabling faster learning. See also Eliaz and Spiegler (2020); Levy, Razin, and Young (2022) for more related work.

[5]Other criteria include the Likelihood Ratio Test (LRT), Akaike information criterion (AIC), Bayesian information criterion (BIC), and cross-validation (Akaike, 1974; Stone, 1977; Schwarz et al., 1978). AIC and BIC approximate the Bayes factor under certain parametric and prior assumptions; cross-validation, though insensitive to the prior, often requires additional regularization to control complexity.

is $y_t = (a_t + b)\omega + \epsilon_t$, where b is the artist's ability, ω is a fixed market demand, and ϵ_t is a noise term with a known distribution. The true values are $b^* = 1$ and $\omega^* = 2$. Effort and market demand are complements: a stronger market incentivizes greater effort. Suppose the artist knows the structure of the revenue function but is uncertain about market demand. She holds a nondegenerate prior over ω and chooses effort each period to maximize expected sales. If she knew her true ability, she could correctly infer demand over time and eventually settle on the optimal effort level, $a^* = 1$. However, the artist has a potentially biased self-perception, assigning probability 1 to $\hat{b} \in \{0, 1, 2\}$, where $\hat{b} = 2$ represents overconfidence and $\hat{b} = 0$ represents underconfidence.[6] Each self-perception $\hat{b} \neq b^*$ gives rise to a misspecified model of how sales are generated: the artist systematically over- or underestimates expected sales. Now, suppose that the artist also considers a competing model that correctly sets $b^* = 1$, and switches to it if it fits the sales data sufficiently better. Are underconfidence and overconfidence equally likely to persist? My results reveal an interesting asymmetry: overconfidence is more robust than underconfidence. This aligns with extensive psychological evidence that overconfidence is generally more prevalent than underconfidence (Svenson, 1981).

Consider first an underconfident artist who believes her ability is low, $\hat{b} = 0$. Underconfidence causes the artist to attribute higher-than-expected sales to strong market demand, leading her to increase effort. To illustrate, suppose the artist starts with the objectively optimal effort $\hat{a}^1 = 1$; based on observed sales, her belief drifts toward $\hat{\omega}^1 = 4$ (see Eq. (1)). This belief induces her to choose $\hat{a}^2 = 2$. Critically, this change in effort partially corrects her overestimation of demand. Because effort and demand are complements, the marginal return to demand increases with effort, allowing the artist to explain sales with a lower demand estimate, $\hat{\omega}^2 = 3$ (see Eq. (1)). This new belief makes $\hat{a}^1$ optimal again, resulting in a negative feedback loop:

$$(\hat{a}^1 + b^*) \cdot \omega^* = (1 + 1) \cdot 2 = (\hat{a}^1 + \hat{b}) \cdot \hat{\omega}^1 = (1 + 0) \cdot \mathbf{4}, \tag{1}$$

$$(\hat{a}^2 + b^*) \cdot \omega^* = (2 + 1) \cdot 2 = (\hat{a}^2 + \hat{b}) \cdot \hat{\omega}^2 = (2 + 0) \cdot \mathbf{3}. \tag{2}$$

The artist's effort perpetually cycles between 1 and 2, with no single belief about demand fully explaining the sales data; i.e., the initial model lacks a self-confirming equilibrium. By contrast, the correctly specified competing model achieves perfect

[6] This assumption captures the idea that individuals often commit fundamental attribution errors and are slower to change self-perceptions than beliefs about external factors (Miller and Ross, 1975).Heidhues et al. (2018); Ba and Gindin (2023) use the same assumption and show that both over- and underconfidence distort demand inferences and lead to inefficient effort choices in the long run.

accuracy asymptotically. Over time, the artist gathers enough evidence to reject her underconfidence and switch to the competing model.

Now, let's turn to the overconfident artist who believes her ability is $\hat{b} = 2$ while also considering the correct competing model. The artist attributes disappointing sales to weak demand and exerts low effort. This low effort further lowers her belief about demand: since the marginal return to demand decreases, she must underestimate demand even more to rationalize poor sales. Such positively reinforcing dynamics eventually drive her belief to $\hat{\omega} = 1$ and effort to $\hat{a} = 0$:

$$(\hat{a} + b^*) \cdot \omega^* = (0 + 1) \cdot 2 = (\hat{a} + \hat{b}) \cdot \hat{\omega} = (0 + 2) \cdot \mathbf{1}. \tag{3}$$

This steady state forms a self-confirming equilibrium: zero effort is optimal given weak demand, and the belief perfectly aligns with the sales data. In the steady state, the initial model with overconfidence and the competing model make equally accurate predictions, and therefore the artist has no reason to switch.

Yet this isn't the whole story. While equilibrium analysis suggests that overconfidence can persist, the agent may switch models before convergence. My dynamic framework examines whether this is possible. I show that for overconfidence to be globally robust—i.e., to persist against the correct model and others—her prior must assign sufficiently high probability to $\hat{\omega} = 1$, with the exact threshold depending on the details of the model-switching rule.

3 Framework

3.1 Basic Setup

Objective Environment. In each period $t = 0, 1, 2, ...$, the agent chooses an action a_t from a finite set $\mathcal{A}$ and observes an outcome y_t drawn from $\mathcal{Y}$ that is either a Euclidean space or a compact subset of it. Both $\mathcal{A}$ and $\mathcal{Y}$ have at least two distinct elements. We equip $\mathcal{Y}$ with a reference measure ν, taken to be counting measure when $\mathcal{Y}$ is finite or countable and Lebesgue measure when $\mathcal{Y}$ is uncountable. Given a_t, the outcome y_t is independently drawn from distribution $Q^*(\cdot|a_t) \in \Delta\mathcal{Y}$. The true data-generating process (DGP), $\{Q^*(\cdot|a)\}_{a\in\mathcal{A}} \in (\Delta\mathcal{Y})^{|\mathcal{A}|}$, remains fixed. At the end of period t, the agent obtains a flow payoff $u_t := u(a_t, y_t)$, where $u : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ is

known.[7] Let $h_t := (a_\tau, y_\tau)_{\tau=0}^{t}$ denote the observable history at the end of period t, with $H_t = (\mathcal{A} \times \mathcal{Y})^{t+1}$ denoting the set of all such histories.

Assumption 1. *For all $a \in \mathcal{A}$: (i) $Q^*(\cdot|a)$ is absolutely continuous w.r.t. ν, and its Radon–Nikodym derivative $q^*(\cdot|a)$ is positive and continuous; (ii) $u(a, \cdot) \in L^1(\mathcal{Y}, \mathbb{R}, Q^*(\cdot|a))$.*[8]

By Assumption 1(i), the true DGP has a positive, continuous density $q^*(\cdot|a)$ for each action. When $\mathcal{Y}$ is discrete, it is a probability mass function; when $\mathcal{Y}$ is continuous, it is a probability density function. Assumption 1(ii) ensures that the objective expected period-t payoff, $\overline{u}_t := \int_{\mathcal{Y}} u(a_t, y)\, q^*(y|a_t)\, \nu(dy)$, is well-defined, and that an optimal action exists.

Subjective Models. The agent relies on subjective models to guide her action choices. Every model, generically denoted by θ, is defined by a finite set of predicted data-generating processes, $\{Q^\theta(\cdot|a, \omega)\}_{a \in \mathcal{A}, \omega \in \Omega^\theta}$, where Ω^θ is a model-specific index set labeling the different DGPs within model θ and is referred to as the parameter space. I place no structure on the elements of Ω^θ beyond finiteness.[9] Without loss of generality, no two distinct parameter values within a model yield the same predicted DGP. The agent only considers models satisfying Assumption 2. The collection of all such models, each being a finite subset of $(\Delta\mathcal{Y})^{|\mathcal{A}|}$, forms the *model universe*, denoted by Θ.

Assumption 2. *For each $a \in \mathcal{A}$ and $\omega \in \Omega^\theta$: (i) $Q^\theta(\cdot|a, \omega)$ is absolutely continuous w.r.t. ν, and its Radon-Nikodym derivative $q^\theta(\cdot|a, \omega)$ is positive and continuous; (ii) $u(a, \cdot) \in L^1(\mathcal{Y}, \mathbb{R}, Q^\theta(\cdot|a, \omega))$; (iii) there exists $r_a \in L^2(\mathcal{Y}, \mathbb{R}, \nu)$ such that r_a is continuous and $\left|\ln \frac{q^*(\cdot|a)}{q^\theta(\cdot|a,\omega)}\right| \leq r_a(\cdot)$ a.s.-$Q^*(\cdot|a)$.*

Assumption 2(i) and (ii) mirror Assumption 1, ensuring well-defined densities and expected payoffs for any model prediction. Assumption 2(iii) bounds the log-likelihood ratio between any prediction and the true DGP, ruling out models whose predictions assign zero probability to histories that occur with positive probability, i.e., surprises. This ensures that the agent can update beliefs using Bayes' rule within any model and apply the Bayes factor rule introduced later to switch models.[10]

[7] The true DGP may also directly enter the payoff, as long as this part of the payoff is unobservable so that the outcome is the only source of information about the true DGP.

[8] $L^p(\mathcal{Y}, \mathbb{R}, \nu)$ denotes the space of all functions $g : \mathcal{Y} \to \mathbb{R}$ s.t. $\int |g(y)|^p \nu(dy) < \infty$.

[9] A finite parameter space ensures that any full-support prior assigns positive probability mass to each prediction in the model. This assumption simplifies the main characterization but can be relaxed for most results (see Online Appendix F.5).

[10] One could extend the framework to allow switching when an observed history has positive probability under one model but zero under another, but the current framework does not treat cases in which all models in consideration assign zero probability to an observed history. Ortoleva (2012) develops a framework that explicitly handles updating after such zero-probability events.

A model in Θ is *correctly specified* if its predictions include the true DGP, i.e., $\exists \omega \in \Omega^\theta$ such that $q^*(\cdot|a) \equiv q^\theta(\cdot|a,\omega), \forall a \in \mathcal{A}$, and *misspecified* otherwise. A model is *larger* if it has a larger parameter space. The smallest correctly specified model, denoted by θ^*, consists only of the true DGP and is referred to as the *true model*. A useful special case (though not the most interesting one) is when data is *exogenous*: the true DGP q^* is an action-independent outcome distribution, and a subjective model is simply a finite set of outcome distributions. In this case, a model is correctly specified if it contains the true distribution and misspecified otherwise.

3.2 The Switcher's Problem

The agent considers a finite set of models, $\Theta^\dagger \subseteq \Theta$. A *dogmatic modeler* only uses a single model, and is a *θ-modeler* when $\Theta^\dagger = \{\theta\}$. My focus is on a *switcher*, who uses one model at a time but may switch between models across periods. The main analysis focuses on the two-model case, $\Theta^\dagger = \{\theta, \theta'\}$, with extensions to multiple competing models discussed in Section 6. A switcher's learning environment is defined by the quadruple $E = (\theta, \theta', \pi_0^\theta, \pi_0^{\theta'})$, where θ is the *initial model*, θ' is the *competing model*, and $\pi_0^\theta \in \Delta\Omega^\theta$ and $\pi_0^{\theta'} \in \Delta\Omega^{\theta'}$ are the agent's priors over the models' parameters. Without loss of generality, all priors have full support. The model chosen in period t is denoted by $m_t \in \Theta^\dagger$, with the initial choice set to $m_0 = \theta$. I now describe the sequence of events in each period $t \geq 0$.

Operating within a model. For $t \geq 1$, the agent first updates her beliefs over parameters within each model using Bayes' rule and history h_{t-1}. This generates two recursive belief processes π_t^θ and $\pi_t^{\theta'}$, where

$$\pi_t^\theta(\omega) := \frac{\pi_{t-1}^\theta(\omega) q^\theta(y_{t-1}|a_{t-1},\omega)}{\sum_{\omega' \in \Omega} \pi_{t-1}^\theta(\omega') q^\theta(y_{t-1}|a_{t-1},\omega')}, \forall \omega \in \Omega^\theta, \tag{4}$$

and $\pi_t^{\theta'}$ is updated analogously.

At $t \geq 0$, the agent chooses an action to maximize the expected flow payoff under her current model m_t and belief $\pi_t^{m_t}$. She follows a pure policy under θ, denoted by f^θ, a selection from the correspondence of myopically optimal actions, $A_M^\theta : \Delta\Omega^\theta \rightrightarrows \mathcal{A}$.[11] The policy under θ' is defined analogously by $f^{\theta'}$. While the agent is assumed to be myopic here, in Section 6 I show that most results extend to a forward-looking agent.

Switching across models. Upon observing y_t, the agent selects the model for the

[11] Actions in $A_M^\theta(\pi_t^\theta)$ maximize the expected flow payoff, $\sum_{\omega \in \Omega^\theta} \pi_t^\theta(\omega) \int_{y \in \mathcal{Y}} q^\theta(y|a,\omega) u(a,y) \nu(dy)$.

next period, m_{t+1}. To guide this decision, she computes the Bayes factor λ_t, which compares how well the two models explain the observed data, h_t. Specifically, it is defined as the ratio of the marginal likelihoods of the data under θ' and θ:

$$\lambda_t \coloneqq \ell_t(\theta')/\ell_t(\theta), \tag{5}$$

where $\ell_t(\theta) \coloneqq \sum_{\omega\in\Omega^\theta} \pi_0^\theta(\omega)\ell_t(\theta,\omega)$ is the marginal likelihood of the data under model θ, and $\ell_t(\theta,\omega) \coloneqq \prod_{\tau=0}^{t} q^\theta(y_\tau|a_\tau,\omega)$ is the likelihood conditional on parameter ω. The marginal likelihood $\ell_t(\theta)$ is the probability of observing $(y_0,\ldots,y_t)$ given $(a_0,\ldots,a_t)$ in model θ. For $t \geq 1$, the agent can compute λ_t recursively:

$$\lambda_t = \lambda_{t-1} \cdot \frac{\sum_{\omega'\in\Omega^{\theta'}} \pi_t^{\theta'}(\omega')\, q^{\theta'}(y_t|a_t,\omega')}{\sum_{\omega\in\Omega^\theta} \pi_t^{\theta}(\omega)\, q^{\theta}(y_t|a_t,\omega)}. \tag{6}$$

That is, the agent updates λ_t by comparing how well each model explains the most recent outcome, weighted by the current posterior.

The agent then compares λ_t to a fixed switching threshold $\alpha \geq 1$. If $m_t = \theta$, the agent switches to $m_{t+1} = \theta'$ when $\lambda_t > \alpha$; if $m_t = \theta'$, the agent switches back to $m_{t+1} = \theta$ when $\lambda_t < 1/\alpha$. When $1/\alpha \leq \lambda_t \leq \alpha$, the evidence is insufficient for a switch, and she retains the current model, $m_{t+1} = m_t$. Thus, the threshold α controls *switching stickiness*, with a larger α requiring stronger evidence to justify switching.[12]

3.3 Persistence and Robustness

Given a learning environment $E = (\theta,\theta',\pi_0^\theta,\pi_0^{\theta'})$, the agent eventually either settles on one model or oscillates forever. I focus on whether the initial model eventually persists.

Definition 1. Model θ *persists against* θ' *at priors* π_0^θ *and* $\pi_0^{\theta'}$ if, given $E = (\theta,\theta',\pi_0^\theta,\pi_0^{\theta'})$, there exists $T \geq 0$ such that, with positive probability, $m_t = \theta$ for all $t \geq T$.[13]

While this definition does not speak to the probability's magnitude or which models are used in the short run, it sets a minimal requirement for long-run survival. Note that persistence is not antisymmetric: it is possible that model θ persists against model θ' while θ' simultaneously persists against θ under the same set of priors. Persistence is also prior-sensitive: a model may persist against a competing model at some priors but not others. This is because priors impact model fit either directly by weighting

[12] Treating α as constant captures switching frictions that are plausibly stable over time. The results are qualitatively unchanged if the agent uses a bounded monotone sequence of thresholds above 1.

[13] For a definition of the underlying probability space, see Appendix A.1.

likelihoods and indirectly by affecting the agent's behavior and the outcome distribution. Finally, persistence is relative: a model may persist against one competing model but not another. Since the choice of competing models and priors is often context-dependent and hard to predict, I introduce notions of *robustness*, i.e., the ability to persist against a wide range of models with varying priors.

The scope of robustness depends on the set of admissible competing models and priors, particularly their distance from the initial model and prior, which defines the allowable step size of switching. I introduce two notions of robustness: *global robustness* permits unlimited step size, while *local robustness* is restricted to minimal step size. Formally, a model θ is globally robust at a given prior if it persists regardless of the competing model θ' and the prior assigned to θ'.

Definition 2 (Global robustness)**.** Model θ is *globally robust at prior* π_0^θ if θ persists against every competing model $\theta' \in \Theta$ at π_0^θ and $\pi_0^{\theta'}$ for every $\pi_0^{\theta'} \in \Delta\Omega^{\theta'}$.

In contrast, local robustness requires that there exist $\epsilon > 0$ such that θ persists against nearby models with nearby priors within the relevant ϵ-neighborhoods. Hence, a locally robust model persists as long as the agent considers small changes to the model; otherwise, some local perturbation will eventually replace it. With models defined as finite sets of DGPs, the distance between models, d, is naturally defined by the Hausdorff distance between their sets of DGPs, where the distance between two DGPs is measured by the maximum Prokhorov distance between their implied outcome distributions across actions. The use of the Hausdorff distance ensures that small perturbations of a misspecified model remain misspecified. I define an ϵ-neighborhood of model θ as $N_\epsilon(\theta) := \{\theta' \in \Theta : d(\theta, \theta') < \epsilon\}$. Since π_0^θ and $\pi_0^{\theta'}$ correspond to distributions over DGPs, their distance can also be measured by the Prokhorov distance. With an abuse of notation, the ϵ-neighborhood of prior π_0^θ within the set of possible priors for θ' is $N_\epsilon^{\theta,\theta'}(\pi_0^\theta) := \left\{\pi_0^{\theta'} \in \Delta\Omega^{\theta'} : d(\pi_0^\theta, \pi_0^{\theta'}) < \epsilon\right\}$. See Appendix A.2 for full definitions.

Definition 3 (Local robustness)**.** Model $\theta \in \Theta$ is *locally robust at prior* π_0^θ if there exists $\epsilon > 0$ such that θ persists against every competing model $\theta' \in N_\epsilon(\theta)$ at priors π_0^θ and $\pi_0^{\theta'}$ for every $\pi_0^{\theta'} \in N_\epsilon^{\theta,\theta'}(\pi_0^\theta)$.

3.4 Discussion of Assumptions

Before presenting the results, I discuss several key assumptions of the framework.

Why switching instead of averaging? Since the agent is already considering multiple models, one might ask why she does not maintain a Bayesian belief over them

and aggregate their predictions. First, the framework already allows for nested models: if the agent starts with a "hypermodel" encompassing multiple sub-models and a prior over them, this effectively serves as her initial model. Second, as Savage (1972) notes, Bayesianism is a reasonable description of decision-making only within "modest little worlds," and it is "utterly ridiculous" to expect people to start with "a model of everything." If people have incomplete models, an important part of learning is the ability to move to a model with different predicted DGPs. This cannot be captured by standard Bayesian updating and necessarily involves a non-Bayesian model-switching process. Third, while the agent updates both models for comparison purposes, combining multiple models to guide decisions is fundamentally different. Doing so requires aggregating predictions across structurally distinct models, which increases cognitive demands, working memory load, or other practical costs.[14] Moreover, some models rest on conceptually incompatible assumptions—e.g., geocentric vs. heliocentric models or liberal vs. conservative worldviews—making it cognitively difficult to reason across them simultaneously.

Why the Bayes factor rule? I assume that the agent evaluates model fit using Bayes factors, a choice that is primarily positive rather than normative.[15] Recent experimental evidence finds that individuals frequently select the best-fitting model as indicated by the Bayes factor (Aina and Schneider, 2025; Barron and Fries, 2024). In addition, the Bayes factor rule has intuitive behavioral interpretations. First, when $\alpha = 1$, it is equivalent to selecting the model with the higher posterior probability of containing the true DGP, under the assumption of a uniform prior. Second, it captures the models' cumulative predictive performance. The agent can be thought of as receiving predictions from two experts, each using a different model, and choosing which to follow based on how well their predictions match outcomes over time. Lastly, the Bayes factor can be easily applied to any models without further parametric assumptions. Due to its Bayesian foundation and ease of use, it has been widely used in recent work on model-based learning and persuasion (Schwartzstein and Sunderam, 2021; Aina, 2025; Galperti, 2019).

[14] Using multiple models is particularly demanding for a forward-looking agent. She must weigh the immediate payoffs of each action under every model, anticipate how the resulting outcome will update her belief across models and affect future payoffs, and aggregate these across models.

[15] In simple settings where the agent receives one-shot information and payoffs depend only on the model choice itself and not posteriors (e.g., if each model prescribes a single action), the optimal switching rule indeed compares the Bayes factor to a fixed threshold. In more complex environments like the one considered here, characterizing the optimal switching rule is challenging and dependent on the decision problem.

An alternative worth considering is the Likelihood Ratio Test (LRT), which compares models based on the ratio of their *maximized* likelihoods. That is, the agent computes $\lambda_t^{\max} := \ell_t^{\max}(\theta')/\ell_t^{\max}(\theta)$, where $\ell_t^{\max}(\theta) := \max_{\omega\in\Omega^\theta} \ell_t(\theta,\omega)$ and $\ell_t^{\max}(\theta') := \max_{\omega'\in\Omega^{\theta'}} \ell_t(\theta',\omega')$. Since $\lambda_t^{\max}$ lacks a recursive structure, the agent must recompute the maximum likelihood estimates using all available data in each period, making it less computationally efficient and less plausible for a boundedly rational agent. More importantly, because the LRT ignores beliefs, it can favor a model that contains a better-fitting DGP even if that DGP was assigned infinitesimal probability, which often leads to unintuitive behavior. For example, in model–expert analogy, the LRT could favor an expert who has consistently underperformed in the past.[16] Section 6 formally compares results under the LRT and the Bayes factor rule.

Why stickiness? I allow the agent to exhibit switching stickiness, captured by $\alpha \geq 1$. Stickiness is well observed in reality and can stem from a variety of causes, such as conservatism, concerns about overreacting to noise, or the cognitive and physical costs associated with model switching. In the statistics literature, Kass and Raftery (1995) suggest a threshold of 20 as the standard for "strong evidence". One important goal of this paper is to examine the implications of stickiness for model persistence.

4 Main Results

4.1 Which Models Can Be Robust?

I first characterize which models can be locally or globally robust for at least one prior. To establish a necessary condition for global robustness, I begin with the case where the competing model is correctly specified. A correctly specified model assigns probability 1 to DGPs that correctly predict the outcome distribution in the limit (Easley and Kiefer, 1988). Therefore, any model that persists against it must also achieve perfect accuracy asymptotically. This implies that the agent converges to a *self-confirming equilibrium* (*SCE*), where she chooses myopically optimal actions based on a *consistent* belief that ensures that the model prediction fully aligns with the true

[16] Suppose that the agent flips a potentially biased coin, with two "experts", θ and θ', providing predictions in every period. Expert θ consistently predicts a tails probability of 2/3, while expert θ' has a uniform prior over the tails probability being either 1/4 or 3/4. If the coin lands on tails twice in a row, the agent should intuitively favor the expert who has consistently assigned a higher probability to tails. Note that expert θ' predicts a tails probability of 1/2 and 5/8 for the first two periods—both lower than 2/3. While the Bayes factor indeed favors θ, the LRT instead favors θ' because its MLE prediction 3/4 is higher than 2/3.

outcome distribution.[17] Note that when data are exogenous, belief consistency requires the model to be correctly specified; with endogenous data, however, a misspecified model may admit an SCE.

Definition 4. A strategy $\sigma \in \Delta\mathcal{A}$ is a *self-confirming equilibrium* (SCE) under model θ if there exists a supporting belief $\pi^\theta \in \Delta\Omega^\theta$ such that: (i) σ is myopically optimal against π^θ, $\sigma \in \Delta A_M^\theta(\pi^\theta)$ and (ii) π^θ is consistent with the true DGP at σ, $q^\theta(\cdot|a,\omega) \equiv q^*(\cdot|a)$ for all $a \in \text{supp}(\sigma)$ and all $\omega \in \text{supp}(\pi)$.

But persisting against a correct model requires more than the existence of an SCE; the SCE must also be reachable and stable. In particular, the agent should, with positive probability, eventually play only the actions in the equilibrium support; if actions outside the support were played infinitely often, the Bayes factor would diverge to infinity, triggering a switch. Since a switcher who adopts θ forever will eventually behave like a θ-modeler, this stability must also hold for a θ-modeler. I term this stability *p-absorbingness*, where "p" indicates that the strategy's support is absorbing *with positive probability*.[18]

Definition 5. Strategy $\sigma \in \Delta\mathcal{A}$ is *p-absorbing* under θ if there exists a full-support prior π_0^θ and some $T \geq 0$ such that, with positive probability, a θ-modeler only plays actions in $\text{supp}\,(\sigma)$ for all $t \geq T$.

I say a model has *perfect asymptotic accuracy* or is *asymptotically accurate* if it admits at least one p-absorbing SCE. Lemma 1 shows that perfect asymptotic accuracy is necessary for a model to persist against a correctly specified model.

Lemma 1. *If model θ persists against a correctly specified model θ' at some priors π_0^θ and $\pi_0^{\theta'}$, then there exists a p-absorbing SCE under θ.*

While this may initially appear to be a weak necessary condition for global robustness—and too strong for local robustness—surprisingly, Theorem 1 shows that if switching exhibits stickiness, perfect asymptotic accuracy is both necessary and sufficient for global and local robustness.

[17] By Definition 4, equilibrium beliefs are restricted to being *unitary* (Fudenberg and Levine, 1993), i.e., a single belief π^θ must rationalize every action in $\text{supp}(\sigma)$. This is needed because, if model θ persists against a correctly specified model, the agent's beliefs must converge. Otherwise, continual belief updating will eventually cause the Bayes factor to exceed the switching threshold almost surely (see Lemma 3 in Appendix A).

[18] P-absorbingness differs from other stability notions in the literature in that it does not require convergence of the action sequence or frequency (Esponda et al., 2021; Fudenberg et al., 2021). See Online Appendix D.1 for an example of a p-absorbing SCE where a θ-modeler's actions never converge, and Online Appendix D.2 for an example of an SCE that fails to be p-absorbing.

Theorem 1. *Suppose $\alpha > 1$. Then the following statements are equivalent:*

(i) Model θ is globally robust for at least one (full-support) prior.

(ii) Model θ is locally robust for at least one (full-support) prior.

(iii) There exists a p-absorbing SCE under model θ.

The implications of Theorem 1 are fourfold. First, it provides a formal learning foundation for asymptotically accurate misspecified models by showing that they can be globally robust, persisting against arbitrary competing models.[19] Second, it reveals that global and local robustness are equivalent when the choice of prior is flexible, though the specific priors supporting each may differ. Thus, if a model is not globally robust, the agent does not need to search far for an alternative: models vulnerable to major paradigm shifts are also susceptible to local changes. Third, together with Lemma 1, Theorem 1 implies that a model that fails to be globally robust cannot persist if the agent considers *any* correctly specified model. Finally, the result holds for all $\alpha > 1$. Since the existence of a p-absorbing SCE depends only on behavior under model θ and is independent of the switching rule, this implies that the set of locally or globally robust models remains unchanged as switching becomes stickier.

Theorem 1 is proved in the order of (i)⇒(ii)⇒(iii)⇒(i). The first step is immediate, as global robustness implies local robustness by definition. The second step shows that local robustness requires perfect asymptotic accuracy. Suppose that model θ does not admit a p-absorbing SCE. While the agent can only consider local perturbations, these can be constructed to achieve higher asymptotic accuracy. I construct θ' as a convex combination of θ and the true DGP while staying close to θ. Since Kullback–Leibler (KL) divergence is convex, θ' yields strictly lower KL divergence than θ given some actions that will be played infinitely often under θ. Though this difference may be small, the Bayes factor diverges to infinity as more outcomes are observed, eventually surpassing α and causing the agent to abandon θ.

It remains to show that perfect asymptotic accuracy implies global robustness for at least one prior. Note that while p-absorbingness ensures that the SCE is reachable for a θ-modeler from some prior, this does not extend to a switcher with the same prior, because actions, beliefs, and model choices are endogenous and interdependent. In particular, the very outcomes that drive a θ-modeler to the SCE can prompt a switch away from θ, making its adoption self-defeating. Example 1 illustrates this dynamic.

[19] Endogenous data is the key for misspecified models to be robust. Note that in an exogenous-data environment, Theorem 1 implies that the sufficient and necessary condition for both local robustness and global robustness is that the model is correctly specified.

Example 1. An agent repeatedly chooses between two tasks, $a_t \in \{a^1, a^2\}$, to maximize output $y_t \in \{0, 1\}$, where 0 represents failure and 1 represents success. The true DGP assigns success rate 0.5 to both tasks, but the agent's model θ assumes that success rates depend on the task and luck $\omega \in \Omega^\theta = \{g, b\}$. Under θ, Task 1 is risky, with success rate 0.5 under good luck (g) and 0.3 under bad luck (b); Task 2 is safe, with a fixed success rate of 0.4. Overall, the agent is pessimistic, as success is less likely under θ than under the true DGP. The agent starts with a uniform prior over luck and chooses Task 1 when she deems good luck more likely, $\pi_t^\theta(g) \geq 0.5$. The competing model θ^* correctly predicts the true success rate.[20] His switching threshold is $\alpha = 1.1$.

Choosing Task 1, a^1, is a p-absorbing SCE under θ, supported by a belief in good luck, $\pi^\theta(g) = 1$. However, while a θ-modeler converges to a^1 with positive probability, θ does not persist against model θ^* at π_0^θ. To see this, note that a failure on Task 1 leads the agent to switch to the safe task a^2, and a success leads to switching to the more optimistic model θ^*. In the first case, the agent *stops* updating belief on luck, which prevents him from returning to a^1. Since θ is incorrectly pessimistic about a^2, the agent eventually switches to model θ^* and enters the second case. Then, switching back to model θ only happens if the agent observes more failures than successes, but such a sequence raises the posterior belief in bad luck, leading the agent back to the safe task a^2, which is not an SCE. In either case, the agent must abandon model θ.

Three factors contribute to the self-defeating result. First, model choice and within-model belief are interdependent, as successes that reinforce a^1 also trigger a switch. Second, the agent's action and model choices are sensitive to early outcomes due to both the prior being distant from the SCE belief and the low switching threshold. Finally, the safe task acts as an absorbing "trap" in θ, preventing further belief updating and blocking a return to the SCE.

The proof of Theorem 1 shows that avoiding this trap is possible if the agent's prior is close to the SCE belief, because it makes action and model choices less sensitive to early outcomes. Hence, perfect asymptotic accuracy is sufficient for global robustness for at least one prior. In the proof, I first derive a stronger property from p-absorbingness: for any $\gamma \in (0, 1)$, there exists a prior such that a θ-modeler keeps playing the SCE and stays near the SCE belief with probability at least γ. Second, I apply Ville's maximal inequality for martingales to show that the probability of the likelihood ratio between the competing model and the true model (i.e., the model consisting only of the true

[20] For simplicity, I assume the agent starts with a uniform prior and takes the true model to be the competing model. Similar issues arise with other priors where the agent is not indifferent between tasks, as well as with competing models arbitrarily close to the initial one.

DGP) staying below α is bounded below by a positive constant. Third, I show that this likelihood ratio approximates the Bayes factor λ_t when beliefs on θ are close to the SCE belief. Together, these steps imply that when γ is close to 1, the probability that the switcher *both* plays the SCE and does not switch is strictly positive.

Corollary 1 further characterizes robust models based on model primitives. Since p-absorbingness is defined for a θ-modeler, it depends only on the primitives of model θ. A simple sufficient condition is that the SCE σ is *quasi-strict*, meaning that any action outside its support is strictly suboptimal given the supporting equilibrium belief π^θ, i.e., $\text{supp}\,(\sigma) = A_M^\theta(\pi^\theta)$. Then, so long as beliefs are near π^θ, actions in $\text{supp}(\sigma)$ are strictly optimal. Since σ is self-confirming, a θ-modeler will, with positive probability, stay near the equilibrium belief and play the SCE forever.[21] Note that any correctly specified model admits a quasi-strict SCE.

Corollary 1. *Suppose that $\alpha > 1$. Then model θ is locally or globally robust for at least one prior if there exists a quasi-strict SCE under θ.*

4.2 When Are Models Robust?

Theorem 1 characterizes which models can be robust but not which priors support them. Identifying these priors is infeasible if there are traps as described in Example 1. To address this, I introduce two regularity conditions that eliminate such traps.

Definition 6. Model θ has *no traps* if:

(i) It is identifiable: $q^\theta(\cdot|a,\omega) \neq q^\theta(\cdot|a,\omega')$ for all distinct $\omega, \omega' \in \Omega^\theta$ and $a \in \mathcal{A}$.

(ii) Any p-absorbing SCE under θ is quasi-strict.

Condition (i) ensures that the DGPs in θ always predict different outcome distributions, ruling out absorbing actions that stop belief-updating. Condition (ii) requires that all p-absorbing SCEs be quasi-strict. When (ii) fails, there exists an action that is optimal given an SCE belief but not self-confirming, which can effectively function as a trap, meaning that once played, the agent cannot revert to the SCE.[22]

[21] The proof of Corollary 1 builds on results from Fudenberg et al. (2021), whose results imply that a *uniformly strict* SCE is uniformly stable in the sense that the agent's actions converge to the equilibrium action with arbitrarily high probability if the agent's prior is sufficiently close to the equilibrium belief. I generalize this to mixed equilibria and show that any quasi-strict mixed SCE is p-absorbing for a myopic agent. In Online Appendix F.4, I show that any uniformly quasi-strict mixed SCE is p-absorbing for a forward-looking agent.

[22] See Online Appendix D.3 for an example of this type of traps.

Under these conditions, Theorem 2 shows that while local robustness under sticky switching is prior-free, global robustness requires priors to be concentrated around the p-absorbing SCEs—a property I refer to as prior tightness. A parameter $\omega \in \Omega^\theta$ is said to be *consistent* under model θ if there exists a p-absorbing SCE under θ supported by the pure belief δ_ω. Formally, given the no-trap conditions, ω is consistent if there exists $\sigma \in \Delta\mathcal{A}$ such that every action $a \in \text{supp}(\sigma)$ is myopically optimal at δ_ω and satisfies $q^\theta(\cdot|a,\omega) \equiv q^*(\cdot|a)$. Let C^θ be the set of all consistent parameters in θ. Note that C^θ is nonempty whenever model θ admits at least one p-absorbing SCE, as identifiability ensures that any SCE supporting belief must be pure. Prior tightness pertains to $\pi_0^\theta(C^\theta)$, the prior probability assigned to the consistent parameters.

Theorem 2. *Suppose that $\alpha > 1$ and model $\theta \in \Theta$ has no traps. Then:*

(i) Model θ is globally robust at prior π_0^θ if and only if $\pi_0^\theta(C^\theta) \geq 1/\alpha$.

(ii) Model θ is locally robust at all (full-support) priors if and only if $C^\theta \neq \emptyset$.

Theorem 2 clarifies the fundamental distinction between local and global robustness: limiting the maximal step size of switching does not expand the set of robust models but allows robust models to persist under a broader range of priors. For global robustness, prior tightness must satisfy a closed-form condition: the prior probability of consistent parameters exceeds $1/\alpha$, $\pi_0^\theta(C^\theta) \geq 1/\alpha$. Thus, prior tightness and switching stickiness are substitutes: when α is close to 1, the prior must be tightly concentrated around C^θ, whereas when α is large, the prior can be more diffuse. In fact, any asymptotically accurate model can be globally robust at a given prior, provided that switching is sufficiently sticky.

The condition of prior tightness can be met in two ways. The agent may start with a strong prior in C^θ, or, if the agent is initially agnostic (e.g., follows an ignorance prior), the parameter space may be small enough so that each parameter value receives a large weight. In the latter case, Theorem 2 highlights a trade-off between model size and robustness. While restricting the number of DGPs potentially reduces asymptotic accuracy, it concentrates the prior, thereby increasing $\pi_0^\theta(C^\theta)$. Hence, the most robust models are asymptotically accurate and simple enough that every parameter in model θ induces a p-absorbing SCE—so that $C^\theta = \Omega^\theta$ and $\pi_0^\theta(C^\theta) = 1$. In this case, global robustness holds *unconditionally*, i.e., at all priors and across all levels of stickiness.

Importantly, Theorem 2 implies that correctly specified models are not necessarily more robust than misspecified ones. It is true that while all correctly specified models are locally robust at all priors and globally robust at some priors or when switching

is sufficiently sticky, only some misspecified models—those with perfect asymptotic accuracy—share this property. However, misspecified models with a smaller parameter space or more SCEs can be globally robust at more diffuse priors or lower levels of switching stickiness. A clean illustration comes from studying unconditional global robustness: under identifiability, the only correctly specified model that satisfies $C^\theta = \Omega^\theta$ is the true model θ^*, since only θ^* has every parameter consistent. By contrast, misspecified models can easily satisfy it if there exist multiple distinct p-absorbing SCEs. In some sense, there is only one way to be robustly correct, but many ways to be robustly wrong. I further illustrate the comparison between correctly specified and misspecified models in the application in Section 5.1.

Prior tightness is crucial for global robustness because it plays a key role in shaping a model's early predictions and ensures a good fit during belief convergence. This is most transparent in the exogenous-data environment, where the agent does not choose actions but passively observes outcomes drawn from the same distribution q^*. There, a parameter is consistent iff it predicts the true DGP, so global robustness requires that the model be correctly specified, with a prior that assigns at least $1/\alpha$ probability to the true DGP. To illustrate, let the competing model be θ^*. By Eq. (5), we have

$$\lambda_t = \lambda_{t-1} \cdot \frac{q^*(y_t)}{\pi_t^\theta(C^\theta)\, q^*(y_t) + \sum_{\omega \notin C^\theta} \pi_t^\theta(\omega)\, q^\theta(y_t|\omega)}. \tag{7}$$

While θ predicts nearly perfectly as $\pi_t^\theta(C^\theta) \to 1$, it fits poorly in early periods if $\pi_0^\theta(C^\theta)$ is too small. The limit of λ_t, which reflects the cumulative relative explanatory power of θ' and θ, depends on the prior odds assigned to the true DGP, $1/\pi_0^\theta(C^\theta)$. A permanent switch occurs if $\pi_0^\theta(C^\theta) < 1/\alpha$. By contrast, local robustness does not require prior tightness, as competing models close to θ with similar priors produce nearly identical predictions, which prevents early switches; thus, asymptotic accuracy alone suffices.

The proof of Theorem 2 generalizes this intuition to endogenous-data environments, where even misspecified models can be globally robust. To prove necessity, I construct a competing model that replaces the initial model almost surely if the prior tightness condition is not met. Unlike the true model used in exogenous-data environments, here the competing model consists of DGPs in C^θ and the true DGP, with the latter assigned a small prior probability $\epsilon > 0$. As $\epsilon \to 0$, λ_t is asymptotically bounded below by $1/\pi_0^\theta(C^\theta)$, yielding the same prior condition. An interesting insight emerges from comparing these cases: in exogenous-data environments, the strongest competing model is a simple and accurate model; in contrast, in endogenous environments, it

may be an extreme and misleading one. Although the competing model constructed here is correctly specified, as ϵ approaches zero, the agent is infinitely more likely to converge to a potentially inefficient SCE associated with C^θ than to an objectively optimal action.

To prove sufficiency, I show that prior tightness allows one to construct a finite sequence of outcomes that bring the agent's posterior closer to an equilibrium belief, where only SCE actions are optimal, while keeping the Bayes factor under α. Complications arise when multiple p-absorbing SCEs exist, as it is uncertain which equilibrium belief the agent will converge to and thus the Bayes factor may have different limits. I show that multiple SCEs relax the prior tightness requirement and it suffices for the total prior probability of C^θ to exceed α.

Finally, Theorem 3 shows that when switching is non-sticky ($\alpha = 1$), all notions of robustness coincide: local and global robustness become equivalent, robustness for some prior is equivalent to robustness for all priors, and these hold only when the model exhibits full prior tightness, $C^\theta = \Omega^\theta$.

Theorem 3. *Suppose that model θ has no traps and $\alpha = 1$. Then model θ is locally or globally robust at any full-support prior π_0^θ if and only if $C^\theta = \Omega^\theta$.*

A key insight is that the set of locally robust models and supporting priors shrinks discontinuously at $\alpha = 1$, which highlights how stickiness helps more misspecified models persist. The discontinuity arises because, for any $\alpha > 1$, the distance between the competing and initial models, as well as their priors, can be made arbitrarily small to keep the Bayes factor below α as the agent converges to the SCE. However, at $\alpha = 1$, there always exists a nearby model that eventually fits slightly better than the initial model if $C^\theta \neq \Omega^\theta$. The proof constructs such a model by preserving most DGPs in θ while slightly improving the accuracy of one DGP associated with some $\omega \in \Omega^\theta \setminus C^\theta$.

5 Applications

5.1 Media Bias and Polarization

In this section, I study a media consumption problem and show how a misspecified model about media bias (Groseclose and Milyo, 2005) can lead to robust polarization, despite no ex ante partisan bias. Moreover, individuals abandon a correctly specified model and adopt this misspecified model permanently with arbitrarily high probability.

An agent chooses between left-wing, centrist, and right-wing media outlets, $a_t \in \mathcal{A} = \{a^L, a^M, a^R\}$, each reporting news stories of two types, $y_t \in \mathcal{Y} = \{l, r\}$. The state of the world $\omega \in \Omega = \{\omega^L, \omega^M, \omega^R\}$ determines the true fraction of l stories. Specifically, $\omega^M = 1/2$, while $\omega^L = 1 - \omega^R = \delta > 1/2$, so an l story signals state ω^L and r signals ω^R. Outlets differ in their reporting biases: a^M reports each story truthfully, a^L misreports r as l with probability $x^L \in (0, 1)$, and a^R misreports l as r with probability $x^R \in (0, 1)$. Our analysis compares two models, a misspecified model $\hat{\theta}$ and a correctly specified model θ. Model $\hat{\theta}$ assumes only ω^L and ω^R are possible (extremism), and takes reporting biases to be $\hat{x}^L \in [0, 1)$ and $\hat{x}^R \in [0, 1)$ (naivety). Model θ recognizes all three states and the actual reporting biases, x^L and x^R. Under both models, the agent strictly prefers the media outlet whose leaning matches the state of the world for consumption benefits not modeled here.[23]

I focus on the interesting case where the true state is ω^M. Suppose that outlets a^L and a^R know this but seek to steer a naive agent toward their preferred states, ω^L and ω^R, respectively. Accordingly, they strategically set $x^L = 2\delta - 1 + 2(1 - \delta)\hat{x}^L > \hat{x}^L$ and $x^R = 2\delta - 1 + 2(1 - \delta)\hat{x}^R > \hat{x}^R$ so that a naive agent underestimates the reporting biases and misinterprets their reported stories as perfect evidence indicating ω^L or ω^R. Under these misreporting strategies, outlet a^L reports a fraction of l stories given by $(1 + x_L)/2$ in state ω^M, which exactly matches the fraction the naive agent expects in state ω^L, i.e., $\delta + (1 - \delta)\hat{x}^L$. A similar equivalence holds for a^R.[24]

Given the outlets' misreporting strategies, a^M is the unique SCE under θ, while a^L and a^R form strict SCEs with distinct supporting beliefs under $\hat{\theta}$. Under the correctly specified model θ, the agent eventually infers the true state and subscribes to the centrist outlet. Under the misspecified model $\hat{\theta}$, by contrast, the agent develops strong political beliefs over time and consumes only media in line with those beliefs, leading to polarization. Proposition 1(i) shows that while $\hat{\theta}$ is globally robust at all priors and switching thresholds, θ is globally robust only at a sufficiently concentrated prior or high switching threshold. The proof directly follows from Theorem 2.

Proposition 1. *(i) Model θ is globally robust at prior π_0^θ if and only if $\pi_0^\theta(\omega^M) \geq 1/\alpha$, while model $\hat{\theta}$ is globally robust at all priors given any $\alpha \geq 1$.*

[23] This preference may arise if the state of the world—specifically, the true fraction of l-stories—directly enters the agent's payoff. That is, we may augment her payoff function to depend on ω. If ω exceeds a threshold, the agent favors a^L; if $1 - \omega$ exceeds a threshold, she favors a^R; if the state is in a moderate range, she favors a^M. See Online Appendix E.2 for concrete examples.

[24] If the outlets deviate from these probabilities, model $\hat{\theta}$ loses perfect asymptotic accuracy and the agent must switch away if the competing model is correctly specified.

(ii) Fix any full-support $\pi_0^\theta, \pi_0^{\hat{\theta}}$ and $\alpha < 1/\pi_0^\theta(\omega^M)$. Given $E = (\theta, \hat{\theta}, \pi_0^\theta, \pi_0^{\hat{\theta}})$, model $\hat{\theta}$ eventually replaces θ with positive probability, i.e., $\exists T \in \mathbb{N}$ s.t. $m_t = \hat{\theta}$ for $t \geq T$. Moreover, this probability converges to 1 as $\pi_0^\theta(\omega^L)$ or $\pi_0^\theta(\omega^R)$ approaches 1.

To illustrate, suppose the agent takes either θ or $\hat{\theta}$ as the initial model and considers the true model, θ^*, as the competing model. Although model θ^* correctly includes ω_M, its average predictions are less accurate than those of θ^* because the predictions associated with ω^L and ω^R are incorrect. Yet the misspecified $\hat{\theta}$ can fit data consistently better than θ^*: the Bayes factor can stay above α regardless of the prior. This occurs, for example, if for the first N periods, the agent happens to see only r stories from a^L and l stories from a^R. Since the data suggests minimal reporting bias, model $\hat{\theta}$ accumulates a growing fit advantage relative to θ^* as N increases—the naive agent becomes increasingly convinced that the partisan outlets are more impartial than they truly are. After enough trust is built, if stories from a^M happen to lean in one direction, she gravitates toward either a^L or a^R—one of the SCEs—while reinforcing her belief about the state. The better fit of $\hat{\theta}$ relative to θ^* persists beyond the first N periods because she ends up in an SCE.

Proposition 1(ii) further shows that if the agent initially adopts the correctly specified model θ and considers the misspecified $\hat{\theta}$ to be the competing model, she abandons θ forever with positive probability whenever $\alpha < 1/\pi_0^\theta(\omega^M)$. This result is a special case of a more general claim: under the no-trap conditions, in any environment $E = (\theta, \hat{\theta}, \pi_0^\theta, \pi_0^{\hat{\theta}})$ where $\hat{\theta}$ is globally robust at all priors and θ is not globally robust at π_0^θ, there exists a finite history that induces a switch from θ to $\hat{\theta}$. Intuitively, as discussed earlier, θ fits worse than the true model θ^*, while $\hat{\theta}$ can consistently fit better than θ^* along sample paths where reported stories initially appear unbiased, so a finite sequence of such observations can trigger a switch from θ to $\hat{\theta}$. After that, $\hat{\theta}$ persists since it is globally robust at all priors. Furthermore, as $\pi_0^\theta(\omega^L)$ or $\pi_0^\theta(\omega^R)$ increases to 1, model θ places a larger weight on incorrect states and delivers a poorer data fit on average, so the probability of a permanent switch to $\hat{\theta}$ converges to one.

5.2 Over- and Underconfidence

This section extends the over- and underconfidence example in Section 2 to broader environments. Each period, an agent chooses a_t from a finite set $\mathcal{A} \subset [\underline{a}, \overline{a}]$ to maximize flow payoff $u(a_t, y_t) = g(a_t, b^*, \omega^*) + \eta_t$, where g is twice continuously differentiable, strictly increasing in ability $b^* \in [\underline{b}, \overline{b}]$ and a fundamental $\omega^* \in [\underline{\omega}, \overline{\omega}]$ (e.g., market demand), and strictly concave in a. Noise η_t follows a known zero-mean distribution.

Action and fundamental are either strict complements or strict substitutes, i.e., $g_{a\omega}$ is either strictly positive or negative across a, b, ω values. Following Heidhues et al. (2018), I assume that ability and the fundamental have opposite effects on optimal effort, so $\text{sgn}(g_{ab}) \neq \text{sgn}(g_{a\omega})$. The agent's model assigns probability 1 to some $\hat{b} \in [\underline{b}, \overline{b}]$. Overconfidence corresponds to $\hat{b} > b^*$, while underconfidence corresponds to $\hat{b} < b^*$. Assume that $\mathcal{A}$ contains actions above and below the objectively optimal action a^*, and that the agent strictly prefers one of them to a^* when she believes her ability is $\underline{b}$ or $\overline{b}$ and the fundamental is ω^*. To avoid trivial nonrobustness, I focus on models with complete parameter spaces: for *any* $a \in \mathcal{A}$, there exists a fundamental value $\Omega^\theta(a) \in \Omega^\theta$ that can perfectly explain observed data, i.e., $g(a, \hat{b}, \Omega^\theta(a)) = g(a, b^*, \omega^*)$. Let $\Theta^M \subset \Theta$ denote the set of all models satisfying this condition.

I assume $\alpha > 1$ and focus on the prior-free local robustness since the interesting asymmetry between over- and underconfidence concerns the induced equilibria rather than the prior. Proposition 2 shows that while overconfidence is always locally robust, underconfidence is only locally robust on a union of disconnected intervals.

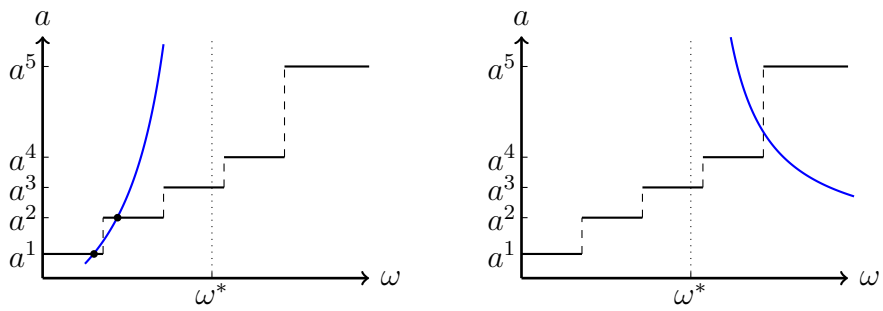
Proposition 2. *There exists a strictly decreasing sequence $b^* = \beta_0 > ... > \beta_K = \underline{b}$ such that any model $\theta \in \Theta^M$ with $\hat{b} \in [\underline{b}, \overline{b}]$ satisfies one of the following conditions.*

(i) When the agent is overconfident, $\hat{b} > b^*$, model θ is locally robust.

(ii) When the agent is underconfident, $\hat{b} < b^*$, model θ is locally robust if $\hat{b} \in (\beta_{2k+1}, \beta_{2k})$ for some $k \in \mathbb{N}$, but not if $\hat{b} \in (\beta_{2k}, \beta_{2k-1})$ for some $k \in \mathbb{N}_+$.

The asymmetry arises because overconfidence always induces a p-absorbing SCE, while underconfidence may not. This stems from their different belief reinforcement dynamics. To illustrate, suppose the fundamental and the action are strict complements ($g_{a\omega} > 0$). The optimal action(s) increase in beliefs about ω, as shown in Fig. 1 by the step curve. For an overconfident agent ($\hat{b} > b^*$), lower a leads to even lower beliefs about ω, as $\Omega^\theta(a') < \Omega^\theta(a'')$ for $a' < a''$. This feedback ensures that the optimal action and inference curves intersect to form at least one strict SCE. For an underconfident agent ($\hat{b} < b^*$), higher a leads to lower beliefs, negatively reinforcing the distortion. Here, the optimal action curve may not intersect the inference curve, leading to no SCE. The agent oscillates between neighboring actions, and her belief does not fully align with data from either.[25] If instead the fundamental and the action are substitutes ($g_{a\omega} < 0$), the orientation of both curves inverts, so Proposition 2 still applies.

[25] The agent converges to a mixed Berk–Nash equilibrium, which extends self-confirming equilibrium by allowing subjective predictions to deviate from the objective outcome distribution in equilibrium (Esponda and Pouzo, 2016).



(a) Overconfidence (b) Underconfidence

Figure 1: Example illustration of equilibria. The step curve is the agent's optimal action correspondence and the blue curve is the agent's inferred fundamental $\Omega^\theta(a)$.

Remark 1. The condition $\text{sgn}(g_{ab}) \neq \text{sgn}(g_{a\omega})$ is sufficient but not necessary. Proposition 2 may still hold in some settings where $\text{sgn}(g_{ab}) = \text{sgn}(g_{a\omega})$, but verifying this requires a case-by-case analysis of belief reinforcement, i.e., whether the inferred fundamental function is comonotone with the optimal action correspondence.[26]

Remark 2. The nonexistence of SCE for a positive measure of $\hat{b}$ below b^* holds for arbitrarily large discrete $\mathcal{A}$. As the number of actions increases, so does the number of unconnected intervals of $\hat{b}$ where θ is not locally robust, with the total measure of such intervals bounded away from zero (see Proposition 4 in Online Appendix E.1). However, since the agent oscillates between actions that are closer together, it can take longer for the Bayes factor to exceed α and trigger a switch. If $\mathcal{A}$ is a continuum, an SCE always exists, even for $\hat{b} < b^*$, and thus over- and underconfidence are symmetric in this dimension. However, characterizing their robustness properties requires extending the framework to continuous actions, which is beyond the scope of this paper.[27]

[26] For example, suppose that $g(a, b, \omega) = ka^n b + a\omega - c(a)$, where $k, n > 0$ and c is strictly increasing. The comparison between over- and underconfidence depends on n. If $n = 1$, then there is no belief reinforcement as $\Omega^\theta(a)$ is independent of a; thus, both over- and underconfidence are locally robust. If $n < 1$, then it can be shown using $g(a, \hat{b}, \Omega^\theta(a)) = g(a, b^*, \omega^*)$ that $\Omega^\theta(a)$ is comonotone with the optimal action correspondence if $\hat{b} > b$, and not comonotone if $\hat{b} < b$. Thus, Proposition 2 still applies. By contrast, if $n > 1$, then the opposite of Proposition 2 holds, i.e., underconfidence is locally robust while overconfidence need not be.

[27] Murooka and Yamamoto (2021) show that a dogmatic modeler with over- or underconfidence converges to an SCE almost surely. However, whether these models can be sustained forever depends on whether actions converge quickly enough to keep the Bayes factor below the switching threshold, which likely depends on the local curvature of the optimal action and inference curves.

6 Extensions

In this section, I discuss how the main results extend when key assumptions are relaxed or modified. Additional extensions, including alternative persistence definitions and infinite parameter spaces, can be found in Online Appendix F.

Alternative Switching Rules. Under the Bayes factor rule, robust models are characterized by asymptotic accuracy and prior tightness. Within the class of Bayes factor rules, the main results remain robust under asymmetric switching thresholds: $\alpha_1 \geq 1$ for switches to θ' and $\alpha_2 \geq 1$ for switches back to θ. However, only the first threshold, α_1, enters the prior tightness condition.[28] Another possible modification concerns the timing of model comparison. Instead of comparing models from the beginning, the agent may wish to wait for T periods to form a more informed posterior and use that as the prior for model comparison (Berger and Pericchi, 1996). This relaxes prior tightness for global robustness and more so for larger T since the agent's posterior converges to an SCE belief. Consequently, asymptotically accurate but misspecified models become even more likely to be robust.[29]

By contrast, the Likelihood Ratio Test (LRT), defined in Section 3.4, leads to starkly different results. Unlike the Bayes factor rule, the LRT relies on maximized likelihoods instead of marginal likelihoods. As shown by Theorem 4 in Online Appendix F.1, for any model $\theta \in \Theta$, there exists a threshold $\overline{\alpha} > 1$ such that if $\alpha \in [1, \overline{\alpha})$, θ cannot be globally robust under any prior. If the outcome space is a continuum or the competing model may contain infinitely many DGPs, then $\overline{\alpha} = \infty$, and hence no model—not even the true model—can be globally robust. The proof constructs larger nesting models that strictly outperform θ over a finite but sufficiently long horizon, ensuring that the agent eventually switches away.

To see the intuition, note that both the Bayes factor rule and the LRT are special cases of a broader class of power-mean likelihood ratio rules, in which the agent computes $\lambda_t^{(\beta)} := \ell_t^{(\beta)}(\theta')/\ell_t^{(\beta)}(\theta)$ for some $\beta \in \mathbb{R}$. Here, $\ell_t^{(\beta)}(\theta) := \left[\sum_{\omega \in \Omega^\theta} \pi_0^\theta(\omega) \cdot \ell_t(\theta, \omega)^\beta\right]^{1/\beta}$ captures the overall *fit* of a model by aggregating the likelihoods of the data across parameters, with higher β placing greater weight on the best-fitting DGP and lower β imposing a penalty on any prior dispersion over DGPs with a worse fit.

[28] This follows from the proofs of Theorems 1 to 3, which either shows that (i) with positive probability the agent never switches or (ii) the Bayes factor exceeds the first switching threshold forever, triggering a permanent switch. In both cases, only the threshold for switching to θ' matters.

[29] An exception occurs when $\alpha = 1$, where Theorem 3 shows that the prior must be fully concentrated at C^θ. Since all model predictions have full support on $\mathcal{Y}$, delaying model comparison does not change this requirement.

As $\beta \to \infty$, the rule converges to the LRT. There, larger models that nest smaller ones always generate weakly higher best fit, so global robustness often becomes impossible. When $\beta = 1$, the Bayes factor rule is recovered, which strikes a balance between best fit and prior dispersion. Accordingly, global robustness is characterized by perfect asymptotic accuracy and prior tightness, with the latter determined by α. As $\beta \to -\infty$, the rule approaches the opposite extreme to the LRT: it evaluates a model's fit based on its *worst* likelihood, which I refer to as the Min-Likelihood Ratio Test (Min-LRT). As shown by Theorem 5 in Online Appendix F.2, under this rule, for any switching threshold $\alpha \geq 1$, only singleton models—those with full prior concentration on a single DGP—that are asymptotically accurate are globally robust.

Multiple Competing Models. The framework extends naturally to multiple competing models, where the agent computes the Bayes factor for each model and switches if any exceeds α. As the number of models, $|\Theta^\dagger|$, increases, global robustness becomes harder to achieve, since it becomes more likely that at least one model will outperform the initial model. If that model turns out to have perfect asymptotic accuracy, the agent may never switch away from it. In fact, if $|\Theta^\dagger| > 2 + \alpha$, even the true model θ^* may fail to be globally robust.[30] A simple fix is to increase switching stickiness: as shown by Theorem 6 in Online Appendix F.3, if $|\Theta^\dagger| < \alpha + 1$, asymptotically accurate misspecified models are still globally robust for some full-support prior.[31]

Forward-Looking Agent. If the agent is forward-looking within each model but not across models, that is, if she maximizes the discounted sum of payoffs under the current model, the main results (Theorems 1 to 3) remain unchanged. While experimentation motives make the p-absorbingness of an SCE harder to achieve, robustness still depends only on the existence of such equilibria. Further details, including a stronger sufficient condition for p-absorbingness, are provided in Online Appendix F.4.

7 Concluding Remarks

This paper develops a theoretical framework for analyzing the robustness of misspecified models when decision-makers are aware of potential model misspecification. By incorporating sticky switching into an active learning framework, I define two notions of model robustness—local and global—and show that they hinge on two properties:

[30] See Example 5 in Online Appendix F.3.

[31] The effect of multiple competing models parallels the multiple comparisons problem in statistics, and increasing stickiness serves a similar role to the Bonferroni correction (Dunn, 1961).

asymptotic accuracy and prior tightness. These results formalize the idea that misspecified models can be persistent if they induce stable self-confirming equilibria and provide a unified framework for comparing robustness across priors, competing models, and switching rules. The framework suggest several directions for future work. Lower switching thresholds make it easier to escape misspecified models but also risk prematurely discarding correctly specified ones due to early noise. An open question is how to optimally set the threshold to balance Type I and Type II errors in the endogenous-data environment. In addition, while persistence ensures that a model may be adopted forever, it does not quantify how likely this adoption is. Future work could study how this depends on key primitives, such as whether the model is correctly specified or misspecified, and features of the learning environment.

References

Aina, C. (2025): "Tailored Stories," *Mimeo*.

Aina, C. and F. H. Schneider (2025): "Weighting Models," *Mimeo*.

Akaike, H. (1974): "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716–723.

Ambuehl, S. and H. C. Thysen (2024): "Choosing Between Causal Interpretations: An Experimental Study," *NHH Dept. of Economics Discussion Paper*.

Ba, C. and A. Gindin (2023): "A Multi-Agent Model of Misspecified Learning with Overconfidence," *Games and Economic Behavior*, 142, 315–338.

Barron, K. and T. Fries (2024): "Narrative persuasion," WZB Discussion Paper SP II 2023-301r, Berlin, additional information: January 2023 (revised April 2024).

Battigalli, P. (1987): "Comportamento Razionale ed Equilibrio nei Giochi e nelle Situazioni Sociali," *Unpublished thesis*.

Berger, J. O. and L. R. Pericchi (1996): "The Intrinsic Bayes Factor for Model Selection and Prediction," *Journal of the American Statistical Association*, 91, 109–122.

Bohren, J. A. (2016): "Informational Herding with Model Misspecification," *Journal of Economic Theory*, 163, 222–247.

Bohren, J. A. and D. N. Hauser (2021): "Learning with Heterogeneous Misspecified Models: Characterization and Robustness," *Econometrica*, 89, 3025–3077.

Cho, I.-K. and K. Kasa (2015): "Learning and Model Validation," *The Review of Economic Studies*, 82, 45–82.

Dunn, O. J. (1961): "Multiple comparisons among means," *Journal of the American statistical association*, 56, 52–64.

Easley, D. and N. M. Kiefer (1988): "Controlling a Stochastic Process with Unknown Parameters," *Econometrica*, 1045–1064.

Eliaz, K. and R. Spiegler (2020): "A Model of Competing Narratives," *American Economic Review*, 110, 3786–3816.

Esponda, I. and D. Pouzo (2016): "Berk–Nash Equilibrium: A Framework for Modeling Agents with Misspecified Models," *Econometrica*, 84, 1093–1130.

Esponda, I., D. Pouzo, and Y. Yamamoto (2021): "Asymptotic Behavior of Bayesian Learners with Misspecified Models," *Journal of Economic Theory*, 195, 105260.

Eyster, E. and M. Rabin (2010): "Naive Herding in Rich-Information Settings," *American Economic Journal: Microeconomics*, 2, 221–43.

Frick, M., R. Iijima, and Y. Ishii (2023): "Belief Convergence under Misspecified Learning: A Martingale Approach," *The Review of Economic Studies*, 90, 781–814.

——— (2024): "Welfare Comparisons for Biased Learning," *American Economic Review*, 114, 1612–1649.

Fudenberg, D. and G. Lanzani (2022): "Which Misspecifications Persist?" *Theoretical Economics*.

Fudenberg, D., G. Lanzani, and P. Strack (2021): "Limit Points of Endogenous Misspecified Learning," *Econometrica*, 89, 1065–1098.

Fudenberg, D. and D. K. Levine (1993): "Self-Confirming Equilibrium," *Econometrica*, 523–545.

Fudenberg, D., G. Romanyuk, and P. Strack (2017): "Active Learning with a Misspecified Prior," *Theoretical Economics*, 12, 1155–1189.

Gagnon-Bartsch, T., M. Rabin, and J. Schwartzstein (2023): "Channeled Attention and Stable Errors," *Mimeo*.

Gagnon-Bartsch, T. and A. Rosato (2024): "Quality is in the eye of the beholder: taste projection in markets with observational learning," *American Economic Review*, 114, 3746–3787.

Galperti, S. (2019): "Persuasion: The Art of Changing Worldviews," *American Economic Review*, 109, 996–1031.

Gilboa, I. and D. Schmeidler (1989): "Maxmin Expected Utility with Non-Unique Prior," *Journal of Mathematical Economics*, 18, 141–153.

Groseclose, T. and J. Milyo (2005): "A Measure of Media Bias," *The Quarterly Journal of Economics*, 120, 1191–1237.

Hansen, L. and T. J. Sargent (2001): "Robust Control and Model Uncertainty," *American Economic Review*, 91, 60–66.

He, K. (2022): "Mislearning from Censored Data: The Gambler's Fallacy and Other Correlational Mistakes in Optimal-Stopping Problems," *Theoretical Economics*, 17, 1269–1312.

He, K. and J. Libgober (2020): "Evolutionarily Stable (Mis) specifications: Theory and Applications," *arXiv preprint arXiv:2012.15007*.

Heidhues, P., B. Kőszegi, and P. Strack (2018): "Unrealistic Expectations and Misguided Learning," *Econometrica*, 86, 1159–1214.

Jehiel, P. (2005): "Analogy-Based Expectation Equilibrium," *Journal of Economic theory*, 123, 81–104.

Jehiel, P. and G. Weber (2025): "Endogenous clustering and analogy-based expectation equilibrium," *Review of Economic Studies*, rdaf054.

Karni, E. and M.-L. Vierø (2013): ""Reverse Bayesianism": A Choice-Based Theory Of Growing Awareness," *American Economic Review*, 103, 2790–2810.

Kass, R. E. and A. E. Raftery (1995): "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Kuhn, T. S. (1962): *The Structure of Scientific Revolutions*, University of Chicago Press.

Lanzani, G. (2025): "Dynamic Concern for Misspecification," *Econometrica*, forthcoming.

Levy, G., R. Razin, and A. Young (2022): "Misspecified Politics and the Recurrence of Populism," *American Economic Review*, 112, 928–962.

Mailath, G. J. and L. Samuelson (2020): "Learning under Diverse World Views: Model-Based Inference," *American Economic Review*, 110, 1464–1501.

Miller, D. T. and M. Ross (1975): "Self-Serving Biases in the Attribution of Causality: Fact or fiction?" *Psychological bulletin*, 82, 213.

Montiel Olea, J. L., P. Ortoleva, M. M. Pai, and A. Prat (2022): "Competing Models," *The Quarterly Journal of Economics*, 137, 2419–2457.

Murooka, T. and Y. Yamamoto (2021): "Misspecified Bayesian Learning by Strategic Players: First-Order Misspecification and Higher-Order Misspecification," *Mimeo*.

——— (2023): "Higher-Order Misspecification and Equilibrium Stability," *Mimeo*.

Nyarko, Y. (1991): "Learning in Mis-specified Models and the Possibility of Cycles," *Journal of Economic Theory*, 55, 416–427.

Ortoleva, P. (2012): "Modeling the Change of Paradigm: Non-Bayesian Reactions to Unexpected News," *American Economic Review*, 102, 2410–36.

Ortoleva, P. and E. Snowberg (2015): "Overconfidence in Political Behavior," *American Economic Review*, 105, 504–35.

Rabin, M. and D. Vayanos (2010): "The Gambler's and Hot-Hand Fallacies: Theory and Applications," *Review of Economic Studies*, 77, 730–778.

Robert, C. P. et al. (2007): *The Bayesian Choice: from Decision-Theoretic Foundations to Computational Implementation*, vol. 2, Springer.

Sargent, T. J. (1999): *The Conquest of American Inflation*, Princeton University Press.

Savage, L. J. (1972): *The Foundations of Statistics*, Courier Corporation.

Schwartzstein, J. and A. Sunderam (2021): "Using Models to Persuade," *American Economic Review*, 111, 276–323.

Schwarz, G. et al. (1978): "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Spiegler, R. (2016): "Bayesian Networks and Boundedly Rational Expectations," *The Quarterly Journal of Economics*, 131, 1243–1290.

——— (2019): "Behavioral Implications of Causal Misperceptions," *Annual Review of Economics*, 12.

——— (2020): "Can Agents with Causal Misperceptions be Systematically Fooled?" *Journal of the European Economic Association*, 18, 583–617.

Stone, M. (1977): "An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion," *Journal of the Royal Statistical Society: Series B (Methodological)*, 39, 44–47.

Svenson, O. (1981): "Are We All Less Risky and More Skillful Than Our Fellow Drivers?" *Acta psychologica*, 47, 143–148.

A Auxiliary Definitions and Results

A.1 Underlying Probability Space

The underlying probability space is constructed as follows. The sample space is defined as $\mathscr{Y} := (\mathcal{Y}^{\infty})^{\mathcal{A}}$, where each element consists of infinite sequences of outcome realizations $(y_{a,0}, y_{a,1}, ...)$ for all actions $a \in \mathcal{A}$. Here, $y_{a,t}$ denotes the outcome if the agent takes a in period t. Independent draws from the true DGP q^* over $\mathcal{Y}$, conditional on actions, induce a product probability measure $\mathbb{P}$ over $\mathscr{Y}$, with $\mathcal{F}$ being the corresponding product sigma-algebra. Let $h := (a_t, y_t)_{t=0}^{\infty}$ denote an infinite history, i.e., an infinite sequence of action-outcome pairs, and define the set of infinite histories as $H := (\mathcal{A} \times \mathcal{Y})^{\infty}$. Given the switching threshold α, the switcher's learning environment $E = (\theta, \theta', \pi_0^{\theta}, \pi_0^{\theta'})$, and the policies $(f^{\theta}, f^{\theta'})$, the probability measure $\mathbb{P}$ induces a measure over H when the agent is a switcher, denoted by $\mathbb{P}_S$. When the agent is a θ-modeler using the prior π_0^{θ} and policy f^{θ}, a different probability measure over H is induced, denoted by $\mathbb{P}_D$. Unless stated otherwise, all probabilistic statements about a switcher are made with respect to $\mathbb{P}_S$, while those concerning a θ-modeler are with respect to $\mathbb{P}_D$.

A.2 Distance Measure for Models

For any two probability measures μ and μ' over metric space $\mathcal{Y}$, the Prokhorov distance is given by

$$d_P(\mu, \mu') := \inf \left\{ \epsilon > 0 | \mu\left(Y\right) \leq \mu'\left(B_{\epsilon}\left(Y\right)\right) + \epsilon \text{ and } \mu'\left(Y\right) \leq \mu\left(B_{\epsilon}\left(Y\right)\right) + \epsilon \text{ for all } Y \subseteq \mathcal{Y} \right\}.$$

The Hausdorff distance between any two sets X and Z is

$$d_H(X, Z) = \max\{\sup_{x \in X} \inf_{z \in Z} \hat{d}(x, z), \sup_{z \in Z} \inf_{x \in X} \hat{d}(x, z)\},$$

where $\hat{d}$ measures the distance between any two elements in X and Z.

For convenience, denote the DGP to which model θ and parameter ω correspond by $Q^{\theta,\omega}$, and the corresponding outcome distribution for action a by $Q_a^{\theta,\omega}$. I define the distance between $Q^{\theta,\omega}$ and $Q^{\theta',\omega'}$ as the maximum Prokhorov distance between the outcome distributions across all actions,

$$d(Q^{\theta,\omega}, Q^{\theta',\omega'}) := \max_{a \in \mathcal{A}} d_P(Q_a^{\theta,\omega}, Q_a^{\theta',\omega'}).$$

The distance between model θ and model θ' is then defined over the sets of DGPs they induce, measured by the Hausdorff metric:

$$d(\theta, \theta') := d_H \left(\{Q^{\theta,\omega}\}_{\omega \in \Omega^\theta}, \{Q^{\theta',\omega'}\}_{\omega' \in \Omega^{\theta'}} \right).$$

All results extend to other commonly used distance measures over probability distributions, including Kullback–Leibler divergence and total variation distance.

A.3 Useful Lemmas

Lemma 2. *Consider any learning environment $E = (\theta, \theta', \pi_0^\theta, \pi_0^{\theta'})$ in which $\theta, \theta' \in \Theta$ and θ' is correctly specified. The ratio $\ell_t(\theta)/\ell_t(\theta')$ a.s. converges to a non-negative finite random variable.*

Proof. Let $\kappa_t = \ell_t(\theta)/\ell_t(\theta')$, then $\kappa_t \geq 0, \forall t$. I now construct the probability measure under which κ_t is a martingale. Given prior $\pi_0^{\theta'}$, denote by $\mathbb{P}_S^{\theta'}$ the probability measure over the set of histories H as implied by model θ'. Formally, for any $\hat{H} \subseteq H$, we have $\mathbb{P}_S^{\theta'}\left(\hat{H}\right) = \sum_{\omega \in \Omega^{\theta'}} \pi_0^{\theta'}(\omega)\, \mathbb{P}_S^{\theta',\omega}\left(\hat{H}\right)$, where $\mathbb{P}_S^{\theta',\omega}$ is the probability measure over H if the true DGP is as described by θ' and ω and the agent is a switcher. Take the conditional expectation of κ_t with respect to $\mathbb{P}_S^{\theta'}$, then we have

$$\begin{aligned}
&\mathbb{E}^{\mathbb{P}_S^{\theta'}}\left(\kappa_t \mid h_{t-1}\right) \\
&= \mathbb{E}^{\mathbb{P}_S^{\theta'}}\left[\frac{\sum_{\omega \in \Omega^\theta} q^\theta\left(y_t|a_t,\omega\right)\pi_t^\theta(\omega)}{\sum_{\omega' \in \Omega^{\theta'}} q^{\theta'}\left(y_t|a_t,\omega'\right)\pi_t^{\theta'}(\omega')} \cdot \kappa_{t-1} \mid h_{t-1}\right] \\
&= \kappa_{t-1} \sum_{\tilde{\omega} \in \Omega^{\theta'}} \pi_t^{\theta'}(\tilde{\omega}) \left[\int_{\mathcal{Y}} \frac{\sum_{\omega \in \Omega^\theta} q^\theta\left(y_t|a_t,\omega\right)\pi_t^\theta(\omega)}{\sum_{\omega' \in \Omega^{\theta'}} q^{\theta'}\left(y_t|a_t,\omega'\right)\pi_t^{\theta'}(\omega')} q^{\theta'}\left(y_t|a_t,\tilde{\omega}\right)\nu\left(dy_t\right)\right] \\
&= \kappa_{t-1} \int_{\mathcal{Y}} \left[\frac{\sum_{\omega \in \Omega^\theta} q^\theta\left(y_t|a_t,\omega\right)\pi_t^\theta(\omega)}{\sum_{\omega' \in \Omega^{\theta'}} q^{\theta'}\left(y_t|a_t,\omega'\right)\pi_t^{\theta'}(\omega')} \left(\sum_{\tilde{\omega} \in \Omega^{\theta'}} q^{\theta'}\left(y_t|a_t,\tilde{\omega}\right)\pi_t^{\theta'}(\tilde{\omega})\right)\right] \nu\left(dy_t\right) \\
&= \kappa_{t-1} \int_{\mathcal{Y}} \left[\sum_{\omega \in \Omega^\theta} q^\theta\left(y_t|a_t,\omega\right)\pi_t^\theta(\omega)\right] \nu\left(dy_t\right) \\
&= \kappa_{t-1} \sum_{\omega \in \Omega^\theta} \left[\int_{\mathcal{Y}} q^\theta\left(y_t|a_t,\omega\right)\nu\left(dy_t\right)\right] \pi_t^\theta(\omega) = \kappa_{t-1}.
\end{aligned}$$

Hence, κ_t is a martingale w.r.t. $\mathbb{P}_S^{\theta'}$. Since κ_t is non-negative, the martingale convergence theorem implies that κ_t converges to a random variable κ and κ is finite almost surely w.r.t. $\mathbb{P}_S^{\theta'}$. Since θ' is correctly specified, there exists a parameter $\omega^* \in \Omega^{\theta'}$

such that $q^*(\cdot|a) \equiv q^{\theta'}(\cdot|a,\omega^*), \forall a \in \mathcal{A}$. It then follows from $\pi_0^{\theta'}(\omega^*) > 0$ that κ_t also converges to κ almost surely w.r.t. $\mathbb{P}_S^{\theta',\omega^*}$, which is the same measure as $\mathbb{P}_S$. □

Lemma 3. *Suppose $\theta \in \Theta$ persists against a correctly specified model $\theta' \in \Theta$ at some full-support priors $\pi_0^\theta, \pi_0^{\theta'}$. Then on paths where m_t eventually equals θ, we have $\lambda_t \xrightarrow{a.s.} \lambda_\infty \leq \alpha$, $\pi_t^{\theta'} \xrightarrow{a.s.} \pi_\infty^{\theta'}$, and $\pi_t^\theta \xrightarrow{a.s.} \pi_\infty^\theta$.*

Proof. It immediately follows from Lemma 2 that $\ell_t(\theta')/\ell_t(\theta) \xrightarrow{\text{a.s.}} \iota \leq \alpha$ on paths where m_t converges to θ. I now show that π_t^θ and $\pi_t^{\theta'}$ also converge almost surely. Given any $\omega \in \Omega^\theta$, write

$$\begin{aligned}\frac{\pi_t^\theta(\omega)}{\pi_0^\theta(\omega)} &= \frac{\prod_{\tau=0}^t q^\theta(y_\tau|a_\tau,\omega)}{\sum_{\omega'\in\Omega^\theta}\prod_{\tau=0}^t q^\theta(y_\tau|a_\tau,\omega')\,\pi_0^\theta(\omega')} \\ &= \frac{\ell_t(\theta')}{\ell_t(\theta)} \cdot \frac{\ell_t(\theta,\omega)}{\ell_t(\theta')},\end{aligned}$$

where the second term $\ell_t(\theta,\omega)/\ell_t(\theta')$ can be seen as the likelihood ratio of a model that consists of a single parameter ω and the competing model θ'. By Lemma 2, $\ell_t(\theta,\omega)/\ell_t(\theta')$ a.s. converges to a random variable that is finite. Consider the paths on which m_t converges to θ. On these paths, both $\ell_t(\theta')/\ell_t(\theta)$ and $\ell_t(\theta,\omega)/\ell_t(\theta')$ converges a.s., which implies that $\pi_t^\theta(\omega)$ a.s. converges to a random variable as well. Since this is true for all $\omega \in \Omega^\theta$, π_t^θ a.s. converges to some limit π_∞^θ on those paths. Analogously, for any $\omega' \in \Omega^{\theta'}$, we can write $\frac{\pi_t^{\theta'}(\omega')}{\pi_0^{\theta'}(\omega')} = \frac{\ell_t(\theta',\omega')}{\ell_t(\theta')}$, which, again by Lemma 2, converges almost surely. □

Lemma 4. *Fix any $\theta, \theta' \in \Theta$, $\omega \in \Omega^\theta, \omega' \in \Omega^{\theta'}$ and any sequence of actions $(a_1, a_2, ...)$. For each infinite history $h \in (\mathcal{A}\times\mathcal{Y})^\infty$ that is generated according to $(a_1, a_2, ...)$ by the true DGP, let*

$$\xi_t(h) = \ln\frac{q^\theta(y_t|a_t,\omega)}{q^{\theta'}(y_t|a_t,\omega')} - \mathbb{E}\left(\ln\frac{q^\theta(y_t|a_t,\omega)}{q^{\theta'}(y_t|a_t,\omega')}|h_{t-1}\right).$$

Then for any fixed $t_0 \geq 1$,

$$\lim_{t\to\infty}(t-t_0+1)^{-1}\sum_{\tau=t_0}^t \xi_\tau(h) = 0, \ \ a.s..$$

Proof. Note that $\xi_t(h)$ is a martingale difference process since $E(\xi_t(h)|h_{t-1}) = 0$. So for any t_0, $\xi_{t_0}^t(h) := \sum_{\tau=t_0}^t (t-\tau+1)^{-1}\xi_\tau(h)$ is also a martingale difference process.

To use the martingale convergence theorem, I now show that $\sup_t \mathbb{E}\left(\left(\xi_{t_0}^t\right)^2\right) < \infty$. Notice that

$$\begin{aligned}
\mathbb{E}\left(\left(\xi_{t_0}^t\right)^2\right) =& \mathbb{E}\left[\left(\sum_{\tau=t_0}^{t}(t-\tau+1)^{-1}\xi_\tau(h)\right)^2\right] \\
\leq& \sum_{\tau=t_0}^{t}(t-\tau+1)^{-2}\mathbb{E}\left[\left(\xi_\tau(h)\right)^2\right] \\
\leq& \sum_{\tau=t_0}^{t}(t-\tau+1)^{-2}\mathbb{E}\left[\left(\ln\frac{q^\theta(y_t|a_t,\omega)}{q^{\theta'}(y_t|a_t,\omega')}\right)^2\right] \\
\leq& 2\sum_{\tau=t_0}^{t}(t-\tau+1)^{-2}\mathbb{E}\left[\left(\ln\frac{q^*(y_t|a_t)}{q^\theta(y_t|a_t,\omega)}\right)^2+\left(\ln\frac{q^*(y_t|a_t)}{q^{\theta'}(y_t|a_t,\omega')}\right)^2\right] \\
\leq& 4\sum_{\tau=t_0}^{t}(t-\tau+1)^{-2}\max_a\mathbb{E}\left[\left(r_a(y_t)\right)^2\right] < \infty,
\end{aligned}$$

where the first inequality follows from the fact that, for any $\tau' > \tau \geq t_0$, $\mathbb{E}\left(\xi_\tau(h)\xi_{\tau'}(h)\right) = \mathbb{E}\left(\mathbb{E}\left(\xi_{\tau'}(h)|h_\tau\right)\xi_\tau(h)\right) = 0$ and the last inequality follows from Assumption 2. Hence, the martingale convergence theorem implies that $\xi_{t_0}^t$ converges to a random variable $\xi_{t_0}^\infty$ almost surely with $\mathbb{E}\left(\left(\xi_{t_0}^\infty\right)^2\right) < \infty$.

Since $\xi_{t_0}^\infty = \lim_{t\to\infty}\sum_{\tau=t_0}^{t}(t-\tau+1)^{-1}\xi_\tau(h)$ is finite a.s., it follows from the Kronecker Lemma that

$$\lim_{t\to\infty}(t-t_0+1)^{-1}\sum_{\tau=t_0}^{t}\xi_\tau(h) = 0, \text{a.s.}.$$

□

B Proofs of Main Results

Unless otherwise stated, I assume throughout that the agent is forward-looking within each model, with a discount factor $\delta \in [0,1)$, but does not anticipate future model switches (see Section 6). This includes the special case where the agent is fully myopic both within and across models, which is the assumption for the main analysis.

For any set of probability distributions $Z \subseteq \Delta S$, define the open ϵ-neighborhood of Z (under the Prokhorov metric d_P) as $B_\epsilon(Z) = \{z \in \Delta S : \inf_{z'\in Z} d_P(z,z') < \epsilon\}$.

For convenience, let $\Omega^\theta(\sigma)$ denote the subset of parameters in model θ such that, for every $a \in \text{supp}(\sigma)$, the distribution $q^\theta(\cdot|a,\omega)$ matches the true DGP $q^*(\cdot|a)$.

B.1 Proof of Lemma 1

By Lemma 3, on paths where θ is eventually forever adopted, beliefs π_t^θ and $\pi_t^{\theta'}$ both converge almost surely. Consider any $\hat{\omega}$ such that with positive probability, m_t eventually equals θ and $\hat{\omega} \in \operatorname{supp}\left(\pi_\infty^\theta\right)$. Let $A^-(\hat{\omega}) := \left\{a \in \mathcal{A} : q^\theta(\cdot|a,\hat{\omega}) \neq q^*(\cdot|a)\right\}$. I now show that every action in $A^-(\hat{\omega})$ is played at most finite times a.s. on the paths where m_t converges to θ and $\hat{\omega} \in \operatorname{supp}\left(\pi_\infty^\theta\right)$. Suppose instead that actions in $A^-(\hat{\omega})$ are played infinitely often. Then there must exist some $\gamma > 0$ such that $\mathbb{E}\ln\frac{q^*(y|a_t)}{q^\theta(y|a_t,\hat{\omega})} > \gamma$ for infinitely many t. Since θ' is correctly specified, there exists a parameter $\omega^* \in \Omega^{\theta'}$ such that $q^*(\cdot|a) \equiv q^{\theta'}(\cdot|a,\omega^*), \forall a \in \mathcal{A}$. Hence, $\mathbb{E}\ln\frac{q^{\theta'}(y|a_t,\omega^*)}{q^\theta(y|a_t,\hat{\omega})} > \gamma$ for infinitely many t. Notice that

$$
\begin{aligned}
\lambda_t = \frac{\ell_t(\theta')}{\ell_t(\theta)} &= \frac{\sum_{\omega'\in\Omega^{\theta'}}\prod_{\tau=0}^t q^{\theta'}(y_\tau|a_\tau,\omega')\,\pi_0^{\theta'}(\omega')}{\sum_{\omega\in\Omega^\theta}\prod_{\tau=0}^t q^\theta(y_\tau|a_\tau,\omega)\,\pi_0^\theta(\omega)} \\
&> \pi_t^\theta(\hat{\omega})\,\frac{\pi_0^{\theta'}(\omega^*)}{\pi_0^\theta(\hat{\omega})}\,\frac{\prod_{\tau=0}^t q^{\theta'}(y_\tau|a_\tau,\omega^*)}{\prod_{\tau=0}^t q^\theta(y_\tau|a_\tau,\hat{\omega})} \\
&= \pi_t^\theta(\hat{\omega})\,\frac{\pi_0^{\theta'}(\omega^*)}{\pi_0^\theta(\hat{\omega})}\exp\left[\sum_{\tau=0}^t 1_{\{a_\tau\in A^-(\hat{\omega})\}}\ln\frac{q^{\theta'}(y_\tau|a_\tau,\omega^*)}{q^\theta(y_\tau|a_\tau,\hat{\omega})}\right],
\end{aligned}
$$

which, by Lemma 4, a.s. increases to infinity as $t \to \infty$, contradicting the assumption that m_t converges to θ. Therefore, on the paths where m_t eventually equals θ, almost surely, there exists T such that $a_t \in \mathcal{A}\backslash \cup_{\hat{\omega}\in\operatorname{supp}(\pi_\infty^\theta)} A^-(\hat{\omega}), \forall t > T$.

Since $q^\theta(\cdot|a,\omega') \equiv q^*(\cdot|a)$ for all $\omega' \in \operatorname{supp}\left(\pi_\infty^\theta\right)$ and all $a \in \mathcal{A}\backslash \cup_{\omega'\in\operatorname{supp}(\pi_\infty^\theta)} A^-(\omega')$, the actions that are played in the limit have no experimentation value and are myopically optimal. Therefore, any strategy that takes support on the limit actions is a self-confirming equilibrium. Fixing a particular value of π_∞^θ that is a limit belief for a positive measure of histories where m_t eventually equals θ, there exists a set of actions $\hat{A} \subseteq A_m^\theta\left(\pi_\infty^\theta\right)$ such that on those histories, the agent only plays actions from this set in the limit. Since m_t eventually converges to θ, it must be true that with positive probability, a θ-modeler who inherits the switcher's prior and policy from the period when the last switch happens also only plays actions from $\hat{A}$ in the limit with positive probability. Therefore, take any strategy σ with $\operatorname{supp}(\sigma) = \hat{A}$, it is a p-absorbing self-confirming equilibrium under θ.

B.2 Proof of Theorem 1

The structure of the proof is as follows. I first show that when $\alpha > 1$, statement (i) is equivalent to statement (iii). The main focus in on the "if" direction since the "only if" direction follows immediately from Lemma 1. I then show that when $\alpha > 1$, statement (ii) implies (iii). Since statement (i) (global robustness) clearly implies statement (ii) (local robustness), it follows that all three statements are equivalent. This completes the proof of Theorem 1.

Step 1. (iii)⇔(i). It directly follows from Lemma 1 that a p-absorbing SCE is necessary for global robustness. I now prove sufficiency.

Pick any competing model $\theta' \in \Theta$ and any full-support prior $\pi_0^{\theta'} \in \Delta\Omega^{\theta'}$. Let $S_t := \ell_t(\theta')/\ell_t(\theta^*)$, then S_t is a martingale with respect to both $\mathbb{P}_D$ and $\mathbb{P}_S$. By Ville's maximal inequality for supermartingales, the probability that S_n is bounded above by a positive constant larger than 1 is bounded away from 0. In particular, for any $\eta \in (1, \alpha)$,

$$\mathbb{P}_D(S_t \leq \eta, \forall t \geq 0) \geq 1 - \frac{\mathbb{E}^{\mathbb{P}_D} S_0}{\eta} = 1 - \frac{1}{\eta}.$$

Note that this inequality holds for any model θ'.

Denote by σ a p-absorbing SCE under θ. Fixing some $\epsilon > 0$, define E as the event that $a_t \in \text{supp}(\sigma)$ and $\pi_t^\theta \in B_\epsilon(\Delta\Omega^\theta(\sigma))$ for all $t \geq 0$. By Lemma 7 in Online Appendix C, p-absorbingness implies that for any $\eta \in (1, \alpha)$, there exists a prior $\pi_0^\theta \in B_\epsilon(\Delta\Omega^\theta(\sigma))$ such that if the agent starts from this prior, $\mathbb{P}_D(E) > 1/\eta$. Therefore,

$$\mathbb{P}_D(E \text{ occurs and } S_t \leq \eta, \forall t \geq 0) \geq \mathbb{P}_D(E) + \mathbb{P}_D(S_t \leq \eta, \forall t \geq 0) - 1 > 0.$$

Denote the histories where E occurs and $S_t \leq \eta, \forall t \geq 0$ by $\hat{H}$. When ϵ is small enough, we have that on $\hat{H}$,

$$\begin{aligned}\lambda_t = \frac{\ell_t(\theta')}{\ell_t(\theta)} &= \frac{\sum_{\omega' \in \Omega^{\theta'}} \pi_0^{\theta'}(\omega') \prod_{\tau=0}^t q^{\theta'}(y_\tau | a_\tau, \omega')}{\sum_{\omega \in \Omega^\theta} \pi_0^\theta(\omega) \prod_{\tau=0}^t q^\theta(y_\tau | a_\tau, \omega)} \\ &< \frac{\sum_{\omega' \in \Omega^{\theta'}} \pi_0^{\theta'}(\omega') \prod_{\tau=0}^t q^{\theta'}(y_\tau | a_\tau, \omega')}{\pi_0^\theta(\Omega^\theta(\sigma)) \prod_{\tau=0}^t q^*(y_\tau | a_\tau)} \leq \frac{\eta}{1-\epsilon} < \alpha\end{aligned}$$

where the first inequality follows from the fact that π_0^θ is full-support and the second inequality follows from the definition of $\hat{H}$. Thus, on $\hat{H}$, the switcher never makes any switch to the competing model θ', i.e., $m_t = \theta, \forall t \geq 0$, and her action choices would be identical to the θ-modeler. Therefore, if we endow the switcher with the same prior

π_0^θ, event $\hat{H}$ also occurs with positive probability under $\mathbb{P}_S$.

Step 2. (ii)⇒(iii). I show that if θ is locally robust at some prior, then it must admit a p-absorbing SCE. Suppose it does not, then a competing model θ' can be constructed as follows. Let θ' have the identical parameter space as θ, i.e., $\Omega^{\theta'} = \Omega^\theta$, and let its predictions be given by $q^{\theta'}(\cdot|a,\omega) = \mu q^\theta(\cdot|a,\omega) + (1-\mu)q^*(\cdot|a)$, for all $a \in \mathcal{A}$ and all $\omega \in \Omega^\theta$, where $\mu \in (0,1)$. For any $\epsilon > 0$, when μ is close enough to 1, we have $\theta' \in N_\epsilon(\theta)$. By the definition of local robustness, there exists $\epsilon > 0$ such that θ persists against θ' at some full-support priors π_0^θ and $\pi_0^{\theta'} = \pi_0^\theta$. Consider any $\hat{\omega} \in \Omega^\theta$ such that $\mathbb{P}_S\left(m_t \to \theta \text{ and } \liminf_{t\to\infty} \pi_t^\theta(\hat{\omega}) > 0\right) > 0$. Let $A^-(\hat{\omega}) := \{a \in \mathcal{A} : q^\theta(\cdot|a,\hat{\omega}) \neq q^*(\cdot|a)\}$. Then every action in $A^-(\hat{\omega})$ is played at most finite times a.s. on the path where m_t eventually equals θ and $\liminf_{t\to\infty} \pi_t^\theta(\hat{\omega}) > 0$. Suppose instead that actions in $A^-(\hat{\omega})$ are played infinitely often. Then there must exist some $\gamma > 0$ such that $\mathbb{E}\ln\frac{q^*(y|a_t)}{q^\theta(y|a_t,\hat{\omega})} > \gamma$ for infinitely many t. So we have

$$\mathbb{E}\ln\frac{q^{\theta'}(y|a_t,\hat{\omega})}{q^\theta(y|a_t,\hat{\omega})} = \mathbb{E}\ln\left(\mu + (1-\mu)\frac{q^*(y|a_t)}{q^\theta(y|a_t,\hat{\omega})}\right) > (1-\mu)\gamma$$

where the inequality follows from Jensen's inequality. Therefore,

$$\begin{aligned}
\lambda_t &= \frac{\sum_{\omega\in\Omega^\theta}\prod_{\tau=0}^t q^{\theta'}(y_\tau|a_\tau,\omega)\,\pi_0^\theta(\omega)}{\sum_{\omega\in\Omega^\theta}\prod_{\tau=0}^t q^{\theta}(y_\tau|a_\tau,\omega)\,\pi_0^\theta(\omega)} \\
&> \pi_t^\theta(\hat{\omega})\,\frac{\pi_0^\theta(\hat{\omega})}{\pi_0^\theta(\hat{\omega})}\frac{\prod_{\tau=0}^t q^{\theta'}(y_\tau|a_t,\hat{\omega})}{\prod_{\tau=0}^t q^{\theta}(y_\tau|a_\tau,\hat{\omega})} \\
&= \pi_t^\theta(\hat{\omega})\exp\left[\sum_{\tau=0}^t 1_{\{a_\tau\in A^-(\hat{\omega})\}}\ln\frac{q^{\theta'}(y_\tau|a_t,\hat{\omega})}{q^{\theta}(y_\tau|a_\tau,\hat{\omega})}\right],
\end{aligned}$$

which, by Lemma 4, almost surely diverges to infinity when m_t converges to θ and $\liminf_{t\to\infty}\pi_t^\theta(\hat{\omega}) > 0$. This implies that, letting $\hat{\Omega}^\theta := \{\omega \in \Omega^\theta : \liminf_{t\to\infty}\pi_t^\theta(\hat{\omega}) > 0\}$, on the paths where m_t eventually equals θ, there almost surely exists T such that $a_t \in \mathcal{A}\backslash\cup_{\hat{\omega}\in\hat{\Omega}^\theta} A^-(\hat{\omega}), \forall t > T$. Since $q^\theta(\cdot|a,\hat{\omega})$ is equal to $q^*(\cdot|a)$ for all $\hat{\omega} \in \hat{\Omega}^\theta$ and all $a \in \mathcal{A}\backslash\cup_{\hat{\omega}\in\hat{\Omega}^\theta} A^-(\hat{\omega})$, the posterior π_t^θ must converge to a limit π_∞^θ. The rest of the arguments are identical to those in the proof of Lemma 1; it follows that θ must admit a p-absorbing SCE.

B.3 Proof of Corollary 1

I show that when the agent is myopic, i.e., discount factor $\delta = 0$, any quasi-strict SCE satisfies a stability property stronger than p-absorbingness, which implies Corollary 1.

Lemma 5. *Suppose the agent is myopic and σ is a quasi-strict SCE with supporting belief $\hat{\pi}$. Then for any $\gamma \in (0,1)$, there exists $\epsilon > 0$ such that starting from any prior $\pi_0^\theta \in B_\epsilon(\hat{\pi})$, the probability that the θ-modeler always plays actions in* $\text{supp}(\sigma)$ *for all periods is strictly larger than γ.*

Proof. Let $\mathbb{P}_D^{\theta,\Omega^\theta(\sigma)}$ denote the probability measure over the set of histories as implied by model θ when the possible DGPs are restricted to $\Omega^\theta(\sigma)$. Formally, for any $\hat{H} \subseteq H$, we have

$$\mathbb{P}_D^{\theta,\Omega^\theta(\sigma)}\left(\hat{H}\right) = \frac{1}{\pi_0^\theta(\Omega^\theta(\sigma))} \sum_{\omega \in \Omega^\theta(\sigma)} \pi_0^\theta(\omega)\, \mathbb{P}_D^{\theta,\omega}\left(\hat{H}\right),$$

where $\mathbb{P}_D^{\theta,\omega}$ is the probability measure over H if the true DGP is as described by θ and ω and the agent is a θ-modeler. If $a_t \in \text{supp}(\sigma)$, then the consistency of the SCE implies that $\mathbb{P}_D^{\theta,\Omega^\theta(\sigma)}(Y_t|a_t) = Q^*(Y_t|a_t)$ for $Y_t \subset \mathcal{Y}$.

Then for every $\omega \in \Omega^\theta \backslash \Omega^\theta(\sigma)$, $\frac{\pi_t^\theta(\omega)}{\pi_t^\theta\left(\Omega^\theta(\sigma)\right)}$ is a non-negative martingale with respect to $\mathbb{P}_D^{\theta,\Omega^\theta(\sigma)}$. It follows that $\frac{\pi_t^\theta\left(\Omega^\theta \backslash \Omega^\theta(\sigma)\right)}{\pi_t^\theta\left(\Omega^\theta(\sigma)\right)}$ is also a non-negative martingale w.r.t. $\mathbb{P}_D^{\theta,\Omega^\theta(\sigma)}$. By Ville's maximal inequality for supermartingales, for any $\eta > 0$,

$$\mathbb{P}_D^{\theta,\Omega^\theta(\sigma)}\left(\frac{\pi_t^\theta\left(\Omega^\theta \backslash \Omega^\theta(\sigma)\right)}{\pi_t^\theta\left(\Omega^\theta(\sigma)\right)} \geq \eta \text{ for some } t\right) < \frac{1}{\eta}\frac{\pi_0^\theta\left(\Omega^\theta \backslash \Omega^\theta(\sigma)\right)}{\pi_0^\theta\left(\Omega^\theta(\sigma)\right)}.$$

Since $\pi_t^\theta\left(\Omega^\theta(\sigma)\right) = 1 - \pi_t^\theta\left(\Omega^\theta \backslash \Omega^\theta(\sigma)\right)$, the above inequality implies that

$$\mathbb{P}_D^{\theta,\Omega^\theta(\sigma)}\left(\pi_t^\theta\left(\Omega^\theta \backslash \Omega^\theta(\sigma)\right) \geq \frac{\eta}{1+\eta} \text{ for some } t\right) < \frac{1}{\eta}\frac{\pi_0^\theta\left(\Omega^\theta \backslash \Omega^\theta(\sigma)\right)}{\pi_0^\theta\left(\Omega^\theta(\sigma)\right)}.$$

If σ is quasi-strict, then $\text{supp}(\sigma) = A_m^\theta(\hat{\pi})$. Since A_M^θ is upper hemicontinuous (Lemma 6 in Online Appendix C), there exists $\tilde{\epsilon} > 0$ small enough such that $\text{supp}(\sigma) \supset A_M^\theta(\pi)$ for all $\pi \in B_{\tilde{\epsilon}}(\hat{\pi})$. Pick some $\epsilon \in (0, \tilde{\epsilon})$ and $\pi_0^\theta \in B_\epsilon(\hat{\pi})$, then $\pi_0^\theta\left(\Omega^\theta \backslash \Omega^\theta(\sigma)\right) < \epsilon$ and $a_0 \in \text{supp}(\sigma)$. Note that the ratio $\frac{\pi_t^\theta(\omega)}{\pi_t^\theta(\omega')}$ remain unchanged throughout all periods such that $a_t \in \text{supp}(\sigma)$ for any $\omega, \omega' \in \Omega^\theta(\sigma)$ since ω and ω' prescribe the same outcome distribution. Hence, if $\pi_t^\theta \notin B_{\tilde{\epsilon}}(\hat{\pi})$ for some $t \geq 0$ and $a_1, ..., a_{t-1} \in \text{supp}(\sigma)$, then there exists t such that $\pi_t^\theta\left(\Omega^\theta \backslash \Omega^\theta(\sigma)\right) \geq \pi_0^\theta\left(\Omega^\theta \backslash \Omega^\theta(\sigma)\right) + \tilde{\epsilon} - \epsilon$. Using the previous

inequality,

$$
\begin{aligned}
& \mathbb{P}_D^{\theta,\Omega^\theta(\sigma)}\left(a_0,...,a_{t-1}\in \text{supp}(\sigma) \text{ and } a_t\notin \text{supp}(\sigma) \text{ for some } t\right) \\
\leq\ & \mathbb{P}_D^{\theta,\Omega^\theta(\sigma)}\left(a_0,...,a_{t-1}\in \text{supp}(\sigma) \text{ and } \pi_t^\theta\notin B_{\tilde{\epsilon}}\left(\hat{\pi}\right) \text{ for some } t\geq 0\right) \\
\leq\ & \mathbb{P}_D^{\theta,\Omega^\theta(\sigma)}\left(\pi_t^\theta\left(\Omega^\theta\backslash\Omega^\theta(\sigma)\right)\geq \pi_0^\theta\left(\Omega^\theta\backslash\Omega^\theta(\sigma)\right)+\tilde{\epsilon}-\epsilon \text{ for some } t\right) \\
<\ & \left(\frac{1}{\pi_0^\theta\left(\Omega^\theta\backslash\Omega^\theta(\sigma)\right)+\tilde{\epsilon}-\epsilon}-1\right)\frac{\pi_0^\theta\left(\Omega^\theta\backslash\Omega^\theta(\sigma)\right)}{\pi_0^\theta\left(\Omega^\theta(\sigma)\right)} \\
<\ & \left(\frac{1}{\tilde{\epsilon}-\epsilon}-1\right)\frac{\epsilon}{1-\epsilon}
\end{aligned}
$$

which converges to 0 as ϵ approaches 0. This implies that for any $\gamma\in(0,1)$ we have $\mathbb{P}_D^{\theta,\Omega^\theta(\sigma)}\left(a_t\in\text{supp}(\sigma),\forall t\geq 0\right)>\gamma$ when ϵ is sufficiently small. Since all parameters in $\Omega^\theta(\sigma)$ prescribe the true outcome distribution whenever actions in $\text{supp}(\sigma)$ are played, $\mathbb{P}_D\left(a_t\in\text{supp}(\sigma),\forall t\geq 0\right)=\mathbb{P}_D^{\theta,\Omega^\theta(\sigma)}\left(a_t\in\text{supp}(\sigma),\forall t\geq 0\right)>\gamma$. □

B.4 Proof of Theorem 2

Part (i). I first prove that global robustness requires prior tightness (necessity) and then prior tightness implies global robustness (sufficiency).

Necessity. Suppose θ is globally robust at prior π_0^θ. By Theorem 1, we know that there must exist a p-absorbing SCE under θ. By identifiability in the no-trap condition, any SCE can only be supported by a pure belief, and hence $C^\theta\neq\emptyset$. Suppose for the sake of contradiction that $\pi_0^\theta(C^\theta)<1/\alpha$. I now construct a competing model such that model θ does not persist against this model at π_0^θ.

Consider a competing model $\theta'\in\Theta$ such that it contains the prediction associated with the parameters in C^θ and the true DGP. In particular, let $\Omega^{\theta'}=C^\theta\cup\{\omega^*\}$ and suppose the predictions of model θ' satisfy that for all $a\in\mathcal{A}$,

$$
q^{\theta'}(\cdot|a,\omega)=\begin{cases} q^\theta(\cdot|a,\omega) & \text{if } \omega\in C^\theta, \\ q^*(\cdot|a) & \text{if } \omega=\omega^*. \end{cases}
$$

In addition, pick some $\epsilon\in(0,1)$ and let the prior $\pi_0^{\theta'}$ be

$$
\pi_0^{\theta'}(\omega)=\begin{cases} (1-\epsilon)\frac{\pi_0^\theta(\omega)}{\pi_0^\theta(C^\theta)} & \text{if } \omega\in C^\theta, \\ \epsilon & \text{if } \omega=\omega^*. \end{cases}
$$

Since θ' is correctly specified, by Lemma 1, on the paths where m_t eventually equals θ, the agent eventually only play actions in the support of an SCE almost surely, and her posterior converges to a supporting belief of the SCE, i.e., $\pi_t^\theta(C^\theta) \xrightarrow{\text{a.s.}} 1$. By construction

$$\ell_t(\theta') = (1-\epsilon)\sum_{\omega\in C^\theta}\frac{\pi_0^\theta(\omega)}{\pi_0^\theta(C^\theta)}\ell_t(\theta,\omega) + \epsilon\ell_t(\theta^*),$$

so we have

$$\frac{\ell_t(\theta')}{\ell_t(\theta)} = (1-\epsilon)\frac{\pi_t^\theta(C^\theta)}{\pi_0^\theta(C^\theta)} + \epsilon\frac{\ell_t(\theta^*)}{\ell_t(\theta)}.$$

The first term almost surely converges to $(1-\epsilon)\frac{1}{\pi_0^\theta(C^\theta)}$. Since $\pi_0^\theta(C^\theta) < 1/\alpha$, there exists ϵ sufficiently small such that $\frac{\ell_t(\theta')}{\ell_t(\theta)} > \alpha$ for sufficiently large t, contradicting the assumption that m_t eventually equals θ.

Sufficiency. Suppose $C^\theta \neq \emptyset$ and $\pi_0^\theta(C^\theta) \geq 1/\alpha$. Pick any competing model θ' and a full-support prior $\pi_0^{\theta'}$. I now show that model θ persists against θ' at the given priors. Define a new probability measure $\hat{\mathbb{P}}$ over the action and outcome histories H such that for any histories $\hat{H} \subset H$,

$$\hat{\mathbb{P}}\left(\hat{H}\right) = \sum_{\omega\in C^\theta}\frac{\pi_0^\theta(\omega)}{\pi_0^\theta(C^\theta)}\mathbb{P}_S^{\theta,\omega}\left(\hat{H}\right),$$

where $\mathbb{P}_S^{\theta,\omega}$ is the probability measure over histories induced by the agent switcher if the true DGP is identical to the DGP prescribed by θ and ω. Define the following process,

$$\hat{\lambda}_t := \frac{1}{\pi_0^\theta(C^\theta)}\frac{\ell_t(\theta')}{\sum_{\omega\in C^\theta}\frac{\pi_0^\theta(\omega)}{\pi_0^\theta(C^\theta)}\ell_t(\theta,\omega)}.$$

Then it is a martingale w.r.t. $\hat{\mathbb{P}}$ with $\mathbb{E}^{\hat{\mathbb{P}}}(\hat{\lambda}_t) = 1/\pi_0^\theta(C^\theta)$. Letting $\eta_t := \pi_0^\theta(C^\theta)\hat{\lambda}_t$, then η_t is also a martingale w.r.t. $\hat{\mathbb{P}}$ with $\mathbb{E}^{\hat{\mathbb{P}}}(\eta_t) = 1$. Since $\mathbb{E}^{\hat{\mathbb{P}}}(\eta_0) = 1$, it must be that $\eta_0 = 1$ almost surely, or there exists $\overline{\eta} < 1$ such that $\eta_0 \leq \overline{\eta}$ with positive probability. Suppose for now that the latter is the case.

By definition, $\hat{\lambda}_t \geq \lambda_t$, where the equality holds only if $\Omega^\theta = C^\theta$. Note that

$$\begin{aligned}
\hat{\mathbb{P}}(\lambda_t \leq \alpha, \forall t) &\geq \hat{\mathbb{P}}(\hat{\lambda}_t \leq \alpha, \forall t) \\
&= \hat{\mathbb{P}}(\eta_t \leq \pi_0^\theta(C^\theta)\alpha, \forall t) \\
&\geq \hat{\mathbb{P}}(\eta_0 \leq \overline{\eta} \text{ and } \eta_t \leq \pi_0^\theta(C^\theta)\alpha, \forall t \geq 2) \\
&\geq \hat{\mathbb{P}}(\eta_0 \leq \overline{\eta}) \cdot \inf_{\eta_0 \leq \overline{\eta}} \hat{\mathbb{P}}(\eta_t \leq \pi_0^\theta(C^\theta)\alpha, \forall t \geq 2 | \eta_0) \\
&\geq \hat{\mathbb{P}}(\eta_0 \leq \overline{\eta}) \cdot \left(1 - \frac{\overline{\eta}}{\pi_0^\theta(C^\theta)\alpha}\right) > 0,
\end{aligned}$$

where the first inequality follows from $\hat{\lambda}_t \geq \lambda_t$, the second inequality follows from $\pi_0^\theta(C^\theta) \geq 1/\alpha$, and the fourth inequality follows from Ville's maximal inequality. If $\eta_0 = 1$ almost surely with respect to $\hat{\mathbb{P}}$, then we only need to consider η_t from $t = 2$ and can apply the same argument as above unless $\eta_2 = 1$ almost surely as well. Iterating this argument, the only remaining case is where $\eta_t = 1$ for all t, but in this case $\hat{\mathbb{P}}(\eta_t \leq \pi_0^\theta(C^\theta)\alpha, \forall t) = 1$.

This implies that there exists $\hat{\omega} \in C^\theta$ such that

$$\mathbb{P}_S^{\theta,\hat{\omega}}(\lambda_t \leq \alpha, \forall t) > 0.$$

Since θ has no traps, it is identifiable and all of its p-absorbing SCEs are quasi-strict. Identifiability implies that $\mathbb{P}_S^{\theta,\hat{\omega}}(\lim_{t\to\infty} \pi_t^\theta(\hat{\omega}) = 1) = 1$. With quasi-strictness, by Lemma 5, there exists $\epsilon > 0$ such that the myopically optimal actions must be in the support of an SCE when $\pi_t^\theta(\hat{\omega}) > 1 - \epsilon$. Since the limit belief is degenerate over a singleton, the myopically optimal action is also dynamically optimal in the limit. Taken together, the no-trap conditions imply that there exists $T > 0$ such that with positive probability (w.r.t. $\mathbb{P}_S^{\theta,\hat{\omega}}$), the agent plays only SCE actions after period T and never switches. Denote the set of such histories by $\hat{H}$. For any $\hat{h} \in \hat{H}$, denote the observable history for the first T periods by $\hat{h}_{T-}$ and the history after the first T periods by $\hat{h}_{T+}$. Since T is finite, by absolute continuity (Assumption 2), for any $\hat{h} \in \hat{H}$, the history $\hat{h}_{T-}$ also occurs with positive probability under the true measure $\mathbb{P}_S$. Conditional on $\hat{h}_{T-}$, since the agent plays only SCE actions on $\hat{H}$ after the first T periods, the two

probability measures $\mathbb{P}_S^{\theta,\hat{\omega}}$ and $\mathbb{P}_S$ over $\hat{H}$ are identical to each other. Therefore,

$$\begin{aligned}\mathbb{P}_S(\hat{H}) &= \sum_{\hat{h}\in\hat{H}} \mathbb{P}_S(\hat{h}_{T-})\mathbb{P}_S(\hat{h}_{T+}|\hat{h}_{T-}) \\ &= \sum_{\hat{h}\in\hat{H}} \mathbb{P}_S(\hat{h}_{T-})\mathbb{P}_S^{\theta,\hat{\omega}}(\hat{h}_{T+}|\hat{h}_{T-}) \\ &\geq \min_{\tilde{h}\in\hat{H}} \frac{\mathbb{P}_S(\tilde{h}_{T-})}{\mathbb{P}_S^{\theta,\hat{\omega}}(\tilde{h}_{T-})}\mathbb{P}_S^{\theta,\hat{\omega}}(\hat{H}) > 0.\end{aligned}$$

This means that with positive probability (under the true probability measure $\mathbb{P}_S$), the agent never switches to θ'. Therefore, model θ persists against θ' at priors π_0^θ and $\pi_0^{\theta'}$.

Part (ii). I first prove that $C^\theta \neq \emptyset$ is a necessary condition for local robustness and then show that it is also sufficient.

Necessity. Suppose θ is locally robust at some full-support prior π_0^θ. It follows from Theorem 1 and identifiability that there exists $\hat{\omega} \in \Omega^\theta$ such that the degenerate belief δ_ω supports a p-absorbing SCE under θ, i.e., $C^\theta \neq \emptyset$.

Sufficiency. Suppose model θ has no traps and $C^\theta \neq \emptyset$. I now show that model θ is locally robust for all full-support priors. Take any $\hat{\omega} \in C^\theta$ and any full-support prior π_0^θ. Consider the probability measure $\mathbb{P}_S^{\theta,\hat{\omega}}$, i.e., the probability measure over infinite histories H induced by the switcher if the true DGP is as described by θ and $\hat{\omega}$. By identifiability and Lemma 4, the posterior π_t^θ converges to $\delta_{\hat{\omega}}$ almost surely under $\mathbb{P}_S^{\theta,\hat{\omega}}$. So for any $\mu > 0$, we can find a set of length-T histories $\hat{H}_{T-1}$ with positive measure where the posterior for model θ enters the μ-neighborhood of $\delta_{\hat{\omega}}$, i.e., $\pi_T^\theta \in B_\mu(\delta_{\hat{\omega}})$. Let μ be small enough so that the posterior $\pi_T^\theta(\hat{\omega}) > 1/\sqrt{\alpha}$. By absolute continuity and the finiteness of T, we know $\hat{H}_{T-1}$ is also realized with positive probability under the true measure $\mathbb{P}_S$.

Next I show that for any $\eta \in (0,1)$, we can choose ϵ to be sufficiently small such that for any $\theta' \in N_\epsilon(\theta)$ and prior $\pi_0^{\theta'} \in N_\epsilon^{\theta,\theta'}(\pi_0^\theta)$, the probability that λ_t never exceeds $\sqrt{\alpha}$ before period T is strictly larger than η. For each $\omega \in \Omega^\theta$, with a slight abuse of notation, denote the set of ϵ-nearby parameters within θ' by $N_\epsilon^{\theta,\theta'}(\omega) := \{\omega' \in \Omega^{\theta'} : d(Q^{\theta,\omega}, Q^{\theta',\omega'}) \leq \epsilon\}$. Let ϵ be sufficiently small such that $N_\epsilon^{\theta,\theta'}(\omega)$ is disjoint across Ω^θ.

By construction we have $\pi_0^{\theta'}(N_\epsilon^{\theta,\theta'}(\omega)) \leq \pi_0^\theta(\omega) + \epsilon$. Hence,

$$\lambda_t = \frac{\ell_t(\theta')}{\ell_t(\theta)} = \frac{\sum_{\omega\in\Omega^\theta}\sum_{\omega'\in N_\epsilon^{\theta,\theta'}(\omega)} \pi_0^{\theta'}(\omega')\prod_{\tau=0}^t q^{\theta'}(y_\tau|a_\tau,\omega')}{\sum_{\omega\in\Omega^\theta}\pi_0^\theta(\omega)\prod_{\tau=0}^t q^\theta(y_\tau|a_\tau,\omega)}$$
$$< \frac{\sum_{\omega\in\Omega^\theta}(\pi_0^\theta(\omega)+\epsilon)\sum_{\omega'\in N_\epsilon^{\theta,\theta'}(\omega)} \mu_0(\omega')\prod_{\tau=0}^t q^{\theta'}(y_\tau|a_\tau,\omega')}{\sum_{\omega\in\Omega^\theta}\pi_0^\theta(\omega)\prod_{\tau=0}^t q^\theta(y_\tau|a_\tau,\omega)}$$

where $\mu_0(\omega') := \frac{\pi_0^{\theta'}(\omega')}{\pi_0^{\theta'}\left(N_\epsilon^{\theta,\theta'}(\omega)\right)}$. We can treat the collection of $N_\epsilon^{\theta,\theta'}(\omega)$ as a new model and μ_0 as the associated prior. This allows us to write the sum of the likelihoods in recursive form,

$$\sum_{\omega'\in N_\epsilon^{\theta,\theta'}(\omega)} \mu_0(\omega')\prod_{\tau=0}^t q^{\theta'}(y_\tau|a_\tau,\omega') = \prod_{\tau=0}^t\left[\sum_{\omega'\in N_\epsilon^{\theta,\theta'}(\omega)} \mu_\tau(\omega')q^{\theta'}(y_\tau|a_\tau,\omega')\right].$$

Let $\hat{Q}_\mu := \sum_{\omega'\in\Omega^{\theta'}}\mu(\omega')Q^{\theta',\omega'}$. Note that for any $\mu \in \Delta(N_\epsilon^{\theta,\theta'}(\omega))$, we have $d(Q^{\theta,\omega},\hat{Q}_\mu) \leq \epsilon$. Therefore, by Lemma 9 in Online Appendix C, for any $r > 0$ and $\gamma < 1$, when ϵ is sufficiently small, the probability that

$$\frac{\sum_{\omega'\in N_\epsilon^{\theta,\theta'}(\omega)} \mu_0(\omega')\prod_{\tau=0}^t q^{\theta'}(y_\tau|a_\tau,\omega')}{\prod_{\tau=0}^t q^\theta(y_\tau|a_\tau,\omega)} \leq (1+r)^{t+1} \tag{8}$$

occurs is larger than γ. Since Ω^θ is finite, this implies that for any $r > 0$ and $\eta < 1$, we can find ϵ sufficiently small such that the probability that Eq. (8) occurs for every $\omega \in \Omega^\theta$ is larger than η. Notice that when Eq. (8) occurs for every $\omega \in \Omega^\theta$,

$$\lambda_t < \max_{\omega\in\Omega^\theta}\left(1+\frac{\epsilon}{\pi_0^\theta(\omega)}\right)(1+r)^{t+1}.$$

Hence, for any $\eta > 0$, we can choose ϵ to be sufficiently small so that the probability that λ_t does not exceed $\sqrt{\alpha}$ for $t = 0, ..., T-1$ is larger than η. Denote the length-$(T+1)$ histories where $\lambda_t \leq \sqrt{\alpha}$ for $t = 0, ..., T-1$ as $\tilde{H}_{T-1}$. Recall that $\hat{H}_{T-1}$ is realized with positive probability. Since the choice of η is arbitrary, we can choose ϵ sufficiently small so that the probability that $\hat{H}_{T-1} \cap \tilde{H}_{T-1}$ is strictly positive.

Finally, note that for any $t > T$, we can write

$$\lambda_t = \lambda_{T-1}\frac{\sum_{\omega'\in\Omega^{\theta'}}\prod_{\tau=T}^t \pi_\tau^{\theta'}(\omega')q^{\theta'}(y_\tau|a_\tau,\omega')}{\sum_{\omega\in\Omega^\theta}\prod_{\tau=T}^t \pi_\tau^\theta(\omega)q^\theta(y_\tau|a_\tau,\omega)} := \lambda_{T-1}\lambda_{T,t}.$$

Recall that on histories $\hat{H}_{T-1} \cap \tilde{H}_{T-1}$ we have $\pi_T^\theta(\hat{\omega}) > 1/\sqrt{\alpha}$, so we can use the same arguments as in part (i) to show that $\mathbb{P}_S(\lambda_{T,t} \leq \sqrt{\alpha}, \forall t > T) > 0$. Since on these histories the agent does not switch before period T and ϵ is small enough such that $\lambda_{T-1} < \sqrt{\alpha}$, we have

$$\begin{aligned}
&\mathbb{P}_S(\lambda_t \leq \alpha, \forall t \geq 0) \\
\geq &\mathbb{P}_S(\hat{H}_{T-1} \cap \tilde{H}_{T-1}) \cdot \mathbb{P}_S(\lambda_{T,t} \leq \sqrt{\alpha}, \forall t \geq T) > 0.
\end{aligned}$$

B.5 Proof of Theorem 3

Note that in the proof of Theorem 2, I prove the sufficiency of prior tightness for global robustness without using the assumption that $\alpha > 1$. When $\alpha = 1$, the prior tightness requirement $\pi_0^\theta(C^\theta) = 1$ is equivalent to $C^\theta = \Omega^\theta$. Therefore, $C^\theta = \Omega^\theta$ is also a sufficient condition for global robustness when $\alpha = 1$. Now it suffices to show that $C^\theta = \Omega^\theta$ is a necessary condition for local robustness when $\alpha = 1$.

Suppose $\theta \in \Theta$ admits at least one p-absorbing SCE and $\pi_0^\theta(C^\theta) < 1$. This implies that there exists $\tilde{\omega} \in \Omega^\theta$ such that $\tilde{\omega} \notin C^\theta$. Consider a local perturbation of model θ, denoted by θ', with the same parameter space $\Omega^{\theta'} = \Omega^\theta$ and prior $\pi_0^{\theta'} = \pi_0^\theta$ but slightly different prediction for $\tilde{\omega}$:

$$q^{\theta'}(\cdot|a,\omega) = \begin{cases} q^\theta(\cdot|a,\omega) & \text{if } \omega \neq \tilde{\omega} \\ \mu q^\theta(\cdot|a,\omega) + (1-\mu) q^*(\cdot|a) & \text{if } \omega = \tilde{\omega} \end{cases}$$

Then for any $\epsilon > 0$, when $\mu \in (0,1)$ is close enough to 1, we have $\theta' \in N_\epsilon(\theta)$. Suppose θ is locally robust and thus persists against θ' for sufficiently small ϵ at priors π_0^θ and $\pi_0^{\theta'}$. Then the Bayes factor satisfies

$$\begin{aligned}
\lambda_t &= \frac{\sum_{\Omega^{\theta'}} \pi_0^{\theta'}(\omega') \prod_{\tau=0}^t q^\theta(y_\tau|a_\tau,\omega')}{\sum_{\Omega^\theta} \pi_0^\theta(\omega) \prod_{\tau=0}^t q^\theta(y_\tau|a_\tau,\omega)} \\
&= \frac{\sum_{\omega \neq \tilde{\omega}} \pi_0^\theta(\omega) \prod_{\tau=0}^t q^\theta(y_\tau|a_\tau,\omega) + \pi_0^\theta(\tilde{\omega}) \prod_{\tau=0}^t q^{\theta'}(y_\tau|a_\tau,\tilde{\omega})}{\sum_{\omega \neq \tilde{\omega}} \pi_0^\theta(\omega) \prod_{\tau=0}^t q^\theta(y_\tau|a_\tau,\omega) + \pi_0^\theta(\tilde{\omega}) \prod_{\tau=0}^t q^\theta(y_\tau|a_\tau,\tilde{\omega})}.
\end{aligned}$$

If θ persists against θ', then there exists $T > 0$ such that $\lambda_t \leq \alpha = 1$ for all $t \geq T$,

which holds if and only if $\frac{\prod_{\tau=0}^{t} q^{\theta'}(y_\tau|a_\tau,\tilde{\omega})}{\prod_{\tau=0}^{t} q^{\theta}(y_\tau|a_\tau,\tilde{\omega})} \leq 1$ for all $t \geq T$. This is further equivalent to

$$\sum_{\tau=0}^{t} \ln \frac{\mu q^{\theta}(y_\tau|a_\tau,\tilde{\omega}) + (1-\mu) q^{*}(y_\tau|a_\tau)}{q^{\theta}(y_\tau|a_\tau,\tilde{\omega})} \leq 0, \forall t \geq T.$$

By concavity of the log function, the above inequality holds only when

$$\sum_{\tau=0}^{t} \ln \frac{q^{\theta}(y_\tau|a_\tau,\tilde{\omega})}{q^{*}(y_\tau|a_\tau)} \geq 0, \forall t \geq T. \tag{9}$$

Note that for any $a \in \mathcal{A}$ such that $q^{\theta}(\cdot|a,\tilde{\omega}) \neq q^{*}(\cdot|a)$,

$$D_{KL}\left(q^{*}(y_\tau|a_\tau) \parallel q^{\theta}(y_\tau|a_\tau,\tilde{\omega})\right) > 0.$$

Therefore, Eq. (9) holds only if there exists $T' \in \mathbb{N}_{+}$ such that $q^{\theta}(\cdot|a_t,\tilde{\omega}) = q^{*}(\cdot|a_t)$ for any $t \geq T'$. This contradicts the assumption that $\tilde{\omega} \notin C^{\theta}$. Hence, θ cannot be locally robust. □

B.6 Proof of Proposition 1

Part (i) is a direct corollary of Theorem 2. I now prove a more general result that implies the first half of Part (ii), i.e., model $\hat{\theta}$ eventually replaces θ with positive probability. The proof for the second half of Part (ii) can be found in Online Appendix E.2.

Proposition 3. *Fix any $E = (\theta, \hat{\theta}, \pi_0^{\theta}, \pi_0^{\hat{\theta}})$ such that both θ and $\hat{\theta}$ satisfy the no-trap condition. Suppose $\hat{\theta}$ is globally robust at all priors and θ is not globally robust at π_0^{θ}. Then there exists some $T \in \mathbb{N}$ such that $m_t = \theta, \forall t \geq T$ with positive probability.*

Proof. It suffices to show that the agent switches to $\hat{\theta}$ at least once with positive probability. It then follows from the fact that $\hat{\theta}$ is globally robust at all priors that $\hat{\theta}$ is eventually adopted forever with positive probability.

Suppose the agent adopts θ forever without switching almost surely. Define a new probability measure $\hat{\mathbb{P}}$ over the action and outcome histories H such that for any histories $\hat{H} \subset H$,

$$\hat{\mathbb{P}}\left(\hat{H}\right) = \sum_{\omega' \in C^{\hat{\theta}}} \pi_0^{\hat{\theta}}(\omega') \mathbb{P}_S^{\hat{\theta},\omega'}\left(\hat{H}\right),$$

where $\mathbb{P}_S^{\hat{\theta},\omega'}$ is the probability measure over histories induced by the agent if the DGP prescribed by $\hat{\theta}$ and ω' is the true DGP. Then $\frac{\sum_{\omega \in C^{\theta}} \pi_0^{\theta}(\omega) \ell_t(\theta,\omega)/\pi_0^{\theta}(C^{\theta})}{\ell_t(\hat{\theta})}$ is a martingale

w.r.t. $\hat{\mathbb{P}}$ with an expectation of 1 at $t = 0$. Hence, for any $\eta > 1$, the probability that $\frac{\sum_{\omega \in C^\theta} \pi_0^\theta(\omega)\ell_t(\theta,\omega)/\pi_0^\theta(C^\theta)}{\ell_t(\hat{\theta})} \leq \eta$ for all t is positive (w.r.t. $\hat{\mathbb{P}}$). On the paths where the model choice eventually equals θ, the agent's posterior π_t^θ almost surely (w.r.t. $\hat{\mathbb{P}}$) converges to some δ_ω where $\omega \in C^\theta$. Taken together, on paths where m_t eventually equals θ, it happens with positive probability (w.r.t. $\hat{\mathbb{P}}$) that $\frac{\sum_{\omega \in C^\theta} \pi_0^\theta(\omega)\ell_t(\theta,\omega)/\pi_0^\theta(C^\theta)}{\ell_t(\hat{\theta})} \leq \eta$ for all t and $\pi_t^\theta \xrightarrow{\text{a.s.}} \delta_\omega$ where $\omega \in C^\theta$. This then implies that for any $\epsilon > 0$, we can construct a finite sequence of outcome realizations $(y_0, ..., y_{t-1})$ such that $\frac{\sum_{\omega \in C^\theta} \pi_0^\theta(\omega)\ell_t(\theta,\omega)/\pi_0^\theta(C^\theta)}{\ell_t(\hat{\theta})} \leq \eta$ for all $t \leq T$ and $\pi_T^\theta \in B_\epsilon(\delta_\omega)$ where $\omega \in C^\theta$. Since T is finite, this sequence of outcomes is realized with positive probability under the true measure $\mathbb{P}_S$. Notice that

$$\frac{\ell_T(\hat{\theta})}{\ell_T(\theta)} = \pi_T^\theta(C^\theta)\frac{\ell_T(\hat{\theta})}{\sum_{\omega \in C^\theta} \pi_0^\theta(\omega)\ell_t(\theta,\omega)} \geq \frac{1-\epsilon}{\eta \pi_0^\theta(C^\theta)},$$

where the right-hand side is strictly larger than α when $\pi_0^\theta(C^\theta) < 1/\alpha$ if ϵ is close enough to 0 and η is close enough to 1. This is a contradiction. Therefore, the agent switches from θ to $\hat{\theta}$ with positive probability. □

B.7 Proof of Proposition 2

Without loss of generality, assume $g_{a\omega} > 0$ and $g_{ab} \leq 0$. Define correspondence $I : [\underline{\omega}, \overline{\omega}] \rightrightarrows [\underline{\omega}, \overline{\omega}]$, such that $I(\omega)$ returns all best-fitting fundamentals at any myopically optimal action against the degenerate belief δ_ω. That is, for any $\hat{\omega} \in I(\omega)$, there exists $\hat{a} \in A_m^\theta(\delta_\omega)$ such that

$$g(\hat{a}, \hat{b}, \hat{\omega}) = g(\hat{a}, b^*, \omega^*).$$

When $\hat{b} > b^*$, $I(\omega) < \omega^*$; when $\hat{b} < b^*$, $I(\omega) > \omega^*$. Fix any $\hat{b}$, there exists a strictly increasing sequence $\{\omega_k\}_{k=0}^K$ with $K \geq 1$, $\omega_0 = \underline{\omega}$, $\omega_K = \overline{\omega}$ such that some action denoted by $a^k \in \mathcal{A}$ is the unique myopically optimal action over (ω_{k-1}, ω_k), $a^k < a^{k+1}$, and both a^k and a^{k+1} are myopically optimal at ω_k if $0 < k < K$. Hence, $I(\omega)$ is a constant function within each (ω_{k-1}, ω_k) and contains two elements if $\omega = \omega_k$ for interior k, which are given by $\lim_{\omega \uparrow \omega_k} I(\omega)$ and $\lim_{\omega \downarrow \omega_k} I(\omega)$. If there exists a self-confirming equilibrium under model θ, then it must be supported by a degenerate belief at ω such that $I(\omega) \ni \{\omega\}$. If $I(\omega) = \{\omega\} \subset (\omega_{k-1}, \omega_k)$ for some k, then a_k is a strict SCE (hence p-absorbing) with the supporting belief δ_ω. For convenience, when $I(\omega) = \{\hat{\omega}\}$, I abuse notation and write $I(\omega) = \hat{\omega}$.

Suppose $\hat{b} > b^*$, then I jumps up discontinuously at all cutoffs $\{\omega_k\}_{1 \leq k \leq K-1}$. By assumption, $\min I(\omega_0) \geq \omega_0$ and $I(\omega_K) < \omega^* < \omega_K$. To find a strict SCE, consider the

following procedure. First, check whether $\max I(\omega_0) > \omega_0$. If not, then $I(\omega_0) = \omega_0$, and a^1 is a strict SCE supported by the belief δ_{ω_0}. If instead $\max I(\omega_0) > \omega_0$, iterate over $k = 1, \ldots, K-1$: (1) Check whether $\min I(\omega_k) < \omega_k$. If so, then a^k is a strict SCE supported by some $\omega \in (\omega_{k-1}, \omega_k)$. (2) If not, then $\max I(\omega_k) > \omega_k$, and the procedure continues to ω_{k+1}. If the final case satisfies $\max I(\omega_{K-1}) > \omega_{K-1}$, then a^K is a strict SCE supported by some $\omega \in (\omega_{K-1}, \omega_K)$. By Corollary 1, model θ is locally robust.

Now suppose $\hat{b} < b^*$, then I jumps down discontinuously at the cutoffs $\{\omega_k\}_{1 \leq k \leq K-1}$. Hence, there exists at most one solution to $I(\omega) = \omega$. When $\hat{b} = b^*$, there exists a unique solution to $I(\omega) = \omega$, i.e., $\omega = \omega^*$. Let $\beta_0 = b^*$. Now suppose there exists an SCE $a^\dagger$ when the agent believes his ability is given by some $\tilde{b} < b^*$. If $\max A_M^\theta(\omega^*) = a^\dagger$, then by the upper-hemicontinuity of A_M^θ, when $\hat{b}$ is lower than but sufficiently close to $\tilde{b}$, there exists some $\hat{\omega} > \omega^*$ such that $g(a^\dagger, \hat{b}, \hat{\omega}) = g(a^\dagger, b^*, \omega^*)$ and $a^\dagger$ is the unique myopically optimal action given belief $\delta_{\hat{\omega}}$. It follows that $a^\dagger$ is a strict SCE. When $\hat{b}$ is sufficiently lower than $\tilde{b}$ such that $a^\dagger \in A_M^\theta(\hat{\omega})$ but $a^\dagger \neq \max A_M^\theta(\hat{\omega})$ for the first time, $a^\dagger$ is still an SCE but no longer strict. Given such $\hat{b}$, if the agent believes his ability is $\hat{b} - \epsilon$ and ϵ is sufficiently small,

$$g(a^\dagger, \hat{b} - \epsilon, \hat{\omega}) < g(a^\dagger, b^*, \omega^*),$$

but for any $a^{\dagger\prime} > a^\dagger$,

$$g(a^{\dagger\prime}, \hat{b} - \epsilon, \hat{\omega}) > g(a^{\dagger\prime}, b^*, \omega^*).$$

Therefore, $I(\omega) = \omega$ admits no solution when the agent's self-perception is $\hat{b} - \epsilon$ when ϵ is sufficiently small. Note that when $\hat{b} = b^*$, the objectively optimal a^* is a SCE, so there exists a strict SCE for any $\hat{b}$ lower than but sufficiently close to β_0. By assumption, there exists $\hat{b} \in (\underline{b}, \overline{b})$ such that a^* is no longer optimal at the inferred fundamental. Thus, letting $\beta_1 = \hat{b}$, there exists no SCE for $\tilde{b}$ lower than but sufficiently close to β_1. Iterating this argument leads to the interval structures described in the proposition. □