Available online at www.sciencedirect.com

ScienceDirect

Journal of Economic Theory 183 (2019) 740–785

JOURNAL OF Economic Theory

www.elsevier.com/locate/jet

Learning and self-confirming long-run biases

P. Battigalli [a,*], A. Francetich [b], G. Lanzani [c], M. Marinacci [a]

[a] *Department of Decision Sciences and IGIER, Università Bocconi, Via Röntgen, 1, 20136 Milano, Italy*
[b] *School of Business, University of Washington Bothell, 18115 Campus Way NE, Bothell, WA 98011, USA*
[c] *Department of Economics, Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02139, USA*

Received 29 March 2017; final version received 13 July 2019; accepted 23 July 2019
Available online 30 July 2019

Abstract

We consider an ambiguity averse, sophisticated decision maker facing a recurrent decision problem where information is generated endogenously. In this context, we study self-confirming actions as the outcome of a process of active experimentation. We provide inter alia a learning foundation for self-confirming equilibrium with model uncertainty (Battigalli et al., 2015), and we analyze the impact of changes in ambiguity attitudes on convergence to self-confirming equilibria. We identify conditions under which the set of self-confirming equilibrium actions is invariant to changes in ambiguity attitudes, and yet ambiguity aversion may affect the dynamics. Indeed, we argue that ambiguity aversion tends to stifle experimentation, increasing the likelihood that the decision maker gets stuck into suboptimal "certainty traps."

JEL classification: C72; D81; D83

Keywords: Learning; Stochastic control; Ambiguity aversion; Self-confirming equilibrium

1. Introduction

We study the dynamic behavior of a decision maker (DM) who faces a recurrent decision problem in which the actions he selects depend on the information endogenously gathered through

* Corresponding author.
E-mail addresses: pierpaolo.battigalli@unibocconi.it (P. Battigalli), aletich@uw.edu (A. Francetich), lanzani@mit.edu (G. Lanzani), massimo.marinacci@unibocconi.it (M. Marinacci).

https://doi.org/10.1016/j.jet.2019.07.009
0022-0531/

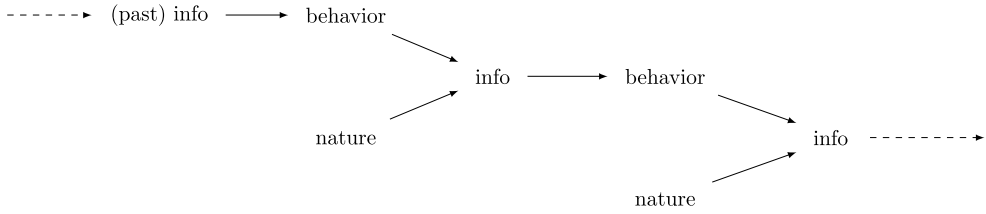

Fig. 1. Timeline.

his past behavior as, for example, in multiarmed bandit problems (cf. Gittins, 1989). We diagram the flow of actions and information in Fig. 1.

Our DM is ambiguity averse, finitely patient, and uncertain about the stochastic process of states of nature. In this setting, there are three crucial elements of our analysis. First, the process of states is governed by an unknown objective probability model (e.g., the composition of an urn). Second, the uncertainty of the DM about the objective model is represented through a subjective probability measure, a belief, which is updated according to information feedback. Each period, the DM evaluates the possible actions (given his updated belief) according to a dynamic version of the smooth ambiguity criterion of Klibanoff et al. (2005), which separates ambiguity attitudes (a personal trait) from the evolving perception of ambiguity, and allows for a Bayesian analysis of learning. Third, the DM uses a rational strategy given his prior belief.

It is essential to understand the meaning of the term "rational" in our setting. An uncertainty averse DM may have *dynamically-inconsistent preferences* (cf. Example 5). While we allow for such reversal of preferences, we assume that the DM is sophisticated in the sense that he formulates a *dynamically-consistent strategy*, that is, a strategy that satisfies the one-deviation (or intrapersonal equilibrium) property: There is no instance where the DM has an incentive to choose an action different from the one prescribed by the given strategy. In a finite-horizon model, this is equivalent to folding-back planning. But we cannot rely on folding back, because we focus on infinite-horizon models to study the limit properties of behavior and beliefs, and to exploit the ensuing stationarity of the dynamic decision problem.

We study how steady-state actions arise from an active experimentation process, providing a novel convergence result. Specifically, we show that the stochastic process of beliefs and actions converges with probability 1 to a random limit action-belief pair. This random limit pair satisfies almost surely the following *self-confirming equilibrium* conditions: The limit action maximizes the one-period value given the limit belief, and the limit belief assigns probability 1 to the set of probability models that are observationally equivalent to the true one given the limit action.

Therefore, even if the DM cares about the future, the limit action-belief pair must be a self-confirming equilibrium of the one-period game repeatedly played against nature. Since the belief may only partially identify the true model (nature's "behavior strategy"), such limit behavior may be very different from the "Nash" (or "rational expectations") equilibrium, in which the DM plays the objective best reply.[1]

Since we assume that the state process is exogenous, that is, the DM's actions cannot influence the probabilities of states in future periods, our framework cannot model long-run interactions with a fixed set of co-players. However, our exogeneity assumption is justified within the scenario

[1] Our definition of self-confirming equilibrium (also called "conjectural equilibrium") is broader than the one of Fudenberg and Levine (1993, 1998). See the discussion in Battigalli et al. (2015).

of large population games. Indeed, our setup can represent the point of view of a DM who plays a game recurrently with other agents independently drawn from large, statistically stable, populations. Hence, the DM recognizes to be unable to influence the evolution of the environment with his actions. The probability models describe the distribution of behaviors in the co-players' populations. Experimentation is valuable to the DM since a better understanding of the correct distribution may allow him to select a better strategy in the following periods. In particular, our setting is consistent with a steady-state learning environment *à la* Fudenberg and He (2018), where individual agents learn through their life, but the population's statistics are constant.

Under this interpretation, we provide a learning foundation for self-confirming equilibrium with model uncertainty (Battigalli et al., 2015, henceforth BCMM). Specifically, the random limit pair corresponds to the "smooth" self-confirming equilibrium concept of BCMM since the limit action is a myopic best response, and the evidence generated by the limit action and the steady-state distribution of opponents' actions confirms the limit belief. BCMM prove that higher ambiguity aversion yields a larger set of self-confirming equilibrium actions. Intuitively, the reason is that a self-confirming equilibrium action is "tested," hence it yields known risks (objective probabilities of consequences), whereas deviations yield unknown risks that are the less attractive the higher the aversion to ambiguity. Since we show that self-confirming equilibrium emerges as the long-run outcome of an active experimentation and learning process, the comparative statics result of BCMM implies that higher ambiguity aversion reduces the predictability of long-run behavior.

We provide special conditions under which the BCMM theorem holds as an invariance result: The set of self-confirming equilibrium actions does not depend on ambiguity attitudes. Nonetheless, ambiguity aversion may still affect the dynamics. Specifically, we argue that ambiguity aversion tends to stifle experimentation, increasing the likelihood that the DM gets stuck into suboptimal "certainty traps." The intuition is as follows. Suppose that the DM can only learn from observing his realized payoffs. The actions perceived as ambiguous, that is, those with uncertain distributions of payoffs, are those from which the DM expects to learn. If instead an action is perceived as unambiguous, the DM expects to have the same belief before and after choosing it, i.e., he does not expect to learn from it. Hence, ambiguity aversion biases the DM toward "exploitation" and against "exploration."

Related literature We point out that there is a formal connection between the concept of SCE and the literature on active learning (or "stochastic control"), and in particular the seminal work by Easley and Kiefer (1988, henceforth EK). The working paper version[2] provides a translation between our setup and the active learning setup. Our paper departs from EK in two fundamental aspects. First, we allow for non-neutral ambiguity attitudes and dynamically inconsistent preferences. Second, EK requires the DM to assign positive subjective probability to (every neighborhood of) the correct model, whereas our sufficient condition for convergence to an SCE allows for misspecified beliefs.[3]

Our definition of self-confirming equilibrium is related to the notion of subjective equilibrium of Kalai and Lehrer (1993 and 1995, henceforth KL). Relatively minor details aside, there are two key differences between the two concepts. First, KL define and analyze subjective equilibrium as the rest point of a process of updated beliefs about the path of play in a supergame. Such

[2] Available as IGIER w.p. 588.

[3] See Section 3. Moreover, the working paper version considers the more general case of non i.i.d. state generating process.

beliefs can be interpreted either as "subjective averages" of probability models, or as subjective Dirac beliefs over probability models. Focusing on such beliefs is without loss of generality under subjective expected utility maximization, but not under non-neutral ambiguity attitudes (see Sections 2.1 and 4). Second, since in KL the set of interacting players is fixed once and for all, their analysis concerns the convergence to a steady state of beliefs about supergame behavior. Our analysis, instead, is consistent with steady-state learning in a population game scenario; thus, we obtain convergence to an equilibrium of the one-period game.

Our results on ambiguity aversion and experimentation are consistent with the findings in Li (2019) and Anderson (2012). Li (2019) characterizes the optimal experimentation strategy under ambiguity aversion in an independent K-armed bandit problem. Aside from focusing on this specific case, the key difference with our paper is that Li (2019) models ambiguity aversion following the two-stage multiple-prior model of Marinacci (2002), while we employ the smooth ambiguity criterion of Klibanoff et al. (2005). As a result, the comparative-statics analysis in Li (2019) considers the impact of changes in the perception of ambiguity, while ours studies the effect of changes in ambiguity attitudes. Moreover, Li (2019) uses a recursive version of the maxmin expected-utility criterion and is thus able to employ standard dynamic programming techniques. Such a recursive representation is precluded in our setting. Unlike Li (2019) and our paper, Anderson (2012) derives the predictions of his model under the implicit assumption that the decision maker can commit ex-ante to any strategy. However, the empirical evidence he presents is consistent with the theoretical predictions of our model.

Outline The paper is structured as follows. Section 2 presents the static and dynamic decision framework, as well as preliminary concepts. Section 3 describes the endogenous information process. Section 4 describes the DM's intertemporal preferences. Section 5 analyzes self-confirming equilibrium, rational strategies, and presents our results on convergence to SCE. Section 6 presents our comparative dynamics results with respect to changes in ambiguity attitudes. Finally, Section 7 briefly relates our analysis to the literature on learning in games and concludes. Proofs are relegated to the appendix. We refer to the working paper version for the complete derivation of the results presented in the examples.

2. Framework

2.1. Static environment

Let S be a finite space of states of nature and let M be a finite outcome space. We consider a control setup where a finite set A of actions (or controls) is available to the DM, and actions and states translate into outcomes through a feedback function $f : A \times S \to M$.[4] We assume that outcomes are observable, while states are not directly observable. Thus, unless the feedback function is injective given the chosen action, inference on the states is partial. The quadruple (A, S, M, f) is the basic structure of the decision problem.

Given a probability measure θ on S, an action a induces a pushforward measure over outcomes via the function $F : A \times \Delta(S) \to \Delta(M)$ defined by:

$$\forall m \in M,\ F(a,\theta)(m) = \sum_{s \in f_a^{-1}(m)} \theta(s),$$

[4] We endow all finite sets with the discrete topology.

where $f_a := f(a, \cdot)$ is the section of f at a. Our first maintained assumption is that the DM is an Expected Utility Maximizer with respect to these (objective) lotteries.

Assumption 1 *(Expected utility on lotteries).* There exists a utility function $u : A \times M \to \mathbb{R}$ such that, for every objective probability measure θ on S and pair of actions $a', a'' \in A$, the DM prefers the (objective) distribution over outcomes induced by a' to the one induced by a'' if and only if:

$$\sum_{m \in M} u(a', m) F(a', \theta)(m) \geq \sum_{m \in M} u(a'', m) F(a'', \theta)(m).$$

Under Assumption 1, we define the expected payoff over outcomes as:

$$R(a, \theta) := \mathbb{E}_{F(a,\theta)}[u_a],$$

where $u_a := u(a, \cdot)$ is the section of u at a.

Let $\Theta \subseteq \Delta(S)$ be a compact set of probability measures on S. These measures, which we call *models*, represent the structural (often physical) information available to the DM, with $\bar{\theta} \in \Theta$ denoting the objectively true model.[5] We identify Θ with a subset of the simplex of dimension $|S| - 1$ and endow it with the Borel σ-algebra $\mathcal{B}(\Theta)$. Incompleteness of information is captured by the non-singleton nature of Θ. Under model uncertainty (cf. Marinacci, 2015), the DM ranks actions according to the *smooth ambiguity* criterion of Klibanoff et al. (2005):

$$\bar{V}(a, \mu) := \phi^{-1}\left(\int_{\Theta} \phi(R(a, \theta))\, \mu(\mathrm{d}\theta)\right), \tag{1}$$

where μ is a prior probability measure on $(\Theta, \mathcal{B}(\Theta))$, and $\phi : \left[\min_{a,\theta} R(a, \theta), \max_{a,\theta} R(a, \theta)\right] \to \mathbb{R}$ is a strictly increasing and continuous function that describes attitudes towards ambiguity.[6] In particular, a concave ϕ captures ambiguity aversion, while a linear ϕ (e.g., the identity function) corresponds to the classical subjective expected utility criterion (Cerreia-Vioglio et al., 2013b)[7]:

$$\bar{V}(a, \mu) = \int_{\Theta} R(a, \theta)\, \mu(\mathrm{d}\theta) = R(a, \theta_\mu),$$

where $\theta_\mu \in \Theta$ is the predictive probability given by $\theta_\mu(E) := \int_\Theta \theta(E)\, \mu(\mathrm{d}\theta)$ for all $E \subseteq S$. Finally, note that:

(i) When the support of μ, $\operatorname{supp} \mu$, is a singleton $\{\theta\}$, criterion (1) reduces to the expected payoff criterion $R(a, \theta)$;

[5] For example, if $S = \{b, g\}$ is the set of possible colors of the ball drawn from an urn of 90 balls that are either black or green, then $\Theta = \{\theta \in \Delta(S) : \theta(b) = i/90 = 1 - \theta(g),\ i \in \{0, ..., 90\}\}$.

[6] See Theorem 6 in Cerreia-Vioglio et al. (2013a) for an axiomatization. In our setting, the domain of ϕ is well defined by finiteness of A and continuity of $R(a, \theta)$ with respect to θ.

[7] To map our decision criterion into theirs, let their space of consequences be $A \times M$ and identify each action a with the act $g(s) = (a, f(a, s))$.

(ii) The limit case of criterion (1) as ambiguity aversion increases is a version of the maximin criterion $\min_{\theta \in \operatorname{supp} \mu} R(a, \theta)$ of Gilboa and Schmeidler (1989); see Proposition 3 in Klibanoff et al. (2005).

The static decision problem can be summarized by:

$$\Gamma = (A, S, M, \Theta, f, u, \phi, \mu). \tag{2}$$

2.2. Dynamic environment

Notation For every finite set Z, we let $Z^t = \prod_{\tau=1}^{t} Z$ and $Z^\infty = \prod_{\tau=1}^{\infty} Z$.[8] We endow the space Z^∞ with the Borel σ-algebra, $\mathcal{B}(Z^\infty)$, corresponding to the product topology on Z^∞; this is the same as the σ-algebra generated by the elementary cylinders $\{z_1\} \times \cdots \times \{z_t\} \times Z^\infty$ (see, e.g., Proposition 1.3 in Folland, 2013). We denote by $z^t = (z_1, ..., z_t) \in Z^t$ both the histories and the elementary cylinders that they identify through the following map:

$$(z_1, ..., z_t) \mapsto \{z_1\} \times \cdots \times \{z_t\} \times Z^\infty .$$

We denote by $z^\infty = (z_1, ..., z_t, ...)$ a generic element of Z^∞.

Environment Given S, let $(S^\infty, \mathcal{B}(S^\infty))$ be the measurable space on which a coordinate state process $(\mathbf{s}_1, \mathbf{s}_2, ...)$ is defined, with $\mathbf{s}_t : S^\infty \to S$ for each t.[9] We will use the less demanding notation $\mathbf{s}^\infty$ for the state process describing the exogenous uncertainty in the decision problem. Its realizations are denoted by $s^\infty \in S^\infty$. Similarly, we write $\mathbf{s}^t = (\mathbf{s}_1, ..., \mathbf{s}_t)$ with realization s^t for finite state processes.

For a generic stochastic process $(\mathbf{z}_1, \mathbf{z}_2, ...)$ defined on $(S^\infty, \mathcal{B}(S^\infty))$, we denote by $\sigma(\mathbf{z}^t)$ the σ-algebra generated by the random variables $\mathbf{z}_1, ..., \mathbf{z}_t$, namely, by the process up to time t; $\sigma(\mathbf{z}^0)$ denotes the trivial σ-algebra.

Finally, for every $\theta \in \Theta$ we define $p_\theta \in \Delta(S^\infty)$ as the unique i.i.d. extension on $\mathcal{B}(S^\infty)$ of the measure given on all elementary cylinders by:

$$p_\theta(s^t \times S^\infty) = \prod_{\tau=1}^{t} \theta(s_\tau)$$

for every t and every s^t in S^t.[10]

Actions and outcomes We describe the DM's choices as a sequence $(a_t) \in A^\infty$ that consists of an action a_t for each period t. At each such t, there is a time-independent feedback function $f : A \times S \to M$, where $f(a_t, s_t)$ is the outcome that the DM observes ex-post (i.e., after the decision) at the end of period t if he chooses action a_t and state s_t obtains.

[8] Unless otherwise stated, it is understood that t is an element of $\mathbb{N}$, the set of natural numbers. We use the terms "time" and "period" interchangeably to refer to t.

[9] We use boldface letters for random variables and normal letters for realizations.

[10] Existence and uniqueness are guaranteed by the Kolmogorov Extension Theorem.

Information feedback In a dynamic setting, the outcome that the DM observes provides feedback about past states, which is a source of "endogenous" (choice dependent) information.[11] Its relevance is peculiar to the dynamic setting and will play a key role in the paper. By selecting action $a_t \in A$ at time t, the DM observes ex-post the outcome $m_t = f(a_t, s_t)$ if state s_t realizes. Thus, a DM who selects action a_t and observes outcome m_t ex-post knows that the realized state s_t belongs to the set $\{s \in S : f(a_t, s) = m_t\} = f_{a_t}^{-1}(m_t)$.

In general, ex-post information about the state is typically endogenous; that is, the partition of the state space S induced by outcomes,

$$\left\{ f_a^{-1}(m) : m \in M \right\},$$

may depend on the choice of action a. If the DM receives the same information about states regardless of his action, namely, if:

$$\forall a, a' \in A,\ \left\{ f_a^{-1}(m) : m \in M \right\} = \left\{ f_{a'}^{-1}(m) : m \in M \right\},$$

we say that feedback satisfies *own-action independence*. In particular, there is *perfect feedback* when the DM observes the realized state s_t ex-post; that is, if f_a is injective for each $a \in A$.

Actions and outcomes are remembered: At each period $t > 1$, the ex-ante endogenous information—that is, the endogenous information gathered prior to the period-t decision—is given by the history of outcomes $m^{t-1} = (m_1, ..., m_{t-1})$ that obtained in the previous periods as a result of the history of actions $a^{t-1} = (a_1, ..., a_{t-1})$ and states $s^{t-1} = (s_1, ..., s_{t-1})$.

Example 1 *(Prelude).* Consider an urn that contains black (B), green (G), and yellow (Y) balls. At each time t, the DM is asked to bet 1 euro on the color of the ball that will be drawn from the urn; therefore, the possible bets are black (b), green (g), and yellow (y). Suppose that the DM is told ex-ante that one-third of the balls are black (and that the only possible colors are B, G, and Y), as in the classical Ellsberg's paradox. That is, the set of posited models is $\Theta = \{\theta \in \Delta(\{B, G, Y\}) : \theta(B) = 1/3\}$. Ex post, after the draw, he only learns the result of his bet, namely, whether or not he wins 1 euro. Here, $S = \{B, G, Y\}$, $A = \{b, g, y\}$, and $M = \{0, 1\}$. The feedback function is described in the following table:

f	B	Y	G
b	1	0	0
y	0	1	0
g	0	0	1

Therefore, we have:

$$f_b^{-1}(1) = \{B\}, \quad f_b^{-1}(0) = \{Y, G\},$$
$$f_y^{-1}(1) = \{Y\}, \quad f_y^{-1}(0) = \{B, G\},$$
$$f_g^{-1}(1) = \{G\}, \quad f_g^{-1}(0) = \{B, Y\}.$$

Note that own-action independence is violated: Ex post, betting on b yields the partition $\{\{B\}, \{Y, G\}\}$ of S, while the bets on y and g respectively yield the partitions $\{\{Y\}, \{B, G\}\}$ *and* $\{\{G\}, \{B, Y\}\}$. ▲

[11] We refer to the working paper version of this paper for a more general setting that separates outcomes and feedback.

Example 2 *(Two-arm bandit).* There are two urns, I and II, with black and green balls. The DM chooses an urn, say k, and wins 1 euro if the ball drawn from urn k is green (G_k, good outcome from urn k) and zero if it is black (B_k, bad outcome from urn k). The outcome for the chosen urn is observed ex-post. Here, $S = \{B_I B_{II}, B_I G_{II}, G_I B_{II}, G_I G_{II}\}$, $A = \{I, II\}$, *and* $M = \{0, 1\}$. The following table describes the feedback function:

f	$B_I B_{II}$	$B_I G_{II}$	$G_I B_{II}$	$G_I G_{II}$
I	0	0	1	1
II	0	1	0	1

Therefore:

$$f_I^{-1}(1) = \{G_I B_{II}, G_I G_{II}\}, \quad f_I^{-1}(0) = \{B_I B_{II}, B_I G_{II}\},$$
$$f_{II}^{-1}(1) = \{B_I G_{II}, G_I G_{II}\}, \quad f_{II}^{-1}(0) = \{B_I B_{II}, G_I B_{II}\}.$$

Own-action independence of feedback is once again violated. ▲

2.3. Strategies and information

Strategies At each period t, the overall ex-ante information available to the DM is given by the histories of actions and outcomes, a^{t-1} and m^{t-1}. The ex-ante information history h_t at time t is given by:

$$h_1 = \left(a^0, m^0\right); \quad \forall t > 1, h_t = \left(a^{t-1}, m^{t-1}\right) = (h_{t-1}, a_{t-1}, m_{t-1}),$$

where $\left(a^0, m^0\right)$ represents null data. Hence, the ex-ante information history space H_{t+1} at the beginning of period $t+1$, determined by information about previous periods, is:

$$H_{t+1} = \left\{(a^t, m^t) \in A^t \times M^t : \exists s^t \in S^t, \forall k \in \{1, ..., t\}, m_k = f(a_k, s_k)\right\}.$$

By definition, $H_1 = \left\{\left(a^0, m^0\right)\right\}$.

Strategies specify an action for each possible information history. Thus, they are modeled as sequences $\alpha = (\alpha_t)$ of time-t strategies, with $\alpha_t : H_t \to A$ for each t. Since $H_1 = \left\{\left(a^0, m^0\right)\right\}$ is a singleton, the first element in the sequence, α_1, prescribes a non-contingent action. Sometimes it is useful to refer to the strategy $\alpha|h_t$ that behaves as specified by h_t at the information sets preceding h_t, and coincides with α elsewhere.

Information and strategies A state process $\mathbf{s}^\infty$ and a strategy $\alpha = (\alpha_t)$ recursively induce an action process $\left(\mathbf{a}_t^\alpha\right)$, an outcome process $\left(\mathbf{m}_t^\alpha\right)$, and an information process $\mathbf{h}^\alpha = \left(\mathbf{h}_t^\alpha\right)$, as follows:

(i) $\mathbf{h}_1^\alpha = \left(a^0, m^0\right)$ and $\mathbf{a}_1^\alpha = \alpha_1\left(a^0, m^0\right)$;
(ii) $\mathbf{m}_1^\alpha = f\left(\mathbf{a}_1^\alpha, \mathbf{s}_1\right)$;
(iii) $\mathbf{h}_{t+1}^\alpha = \left(\mathbf{h}_t^\alpha, \mathbf{a}_t^\alpha, \mathbf{m}_t^\alpha\right)$, $\mathbf{a}_{t+1}^\alpha = \alpha_{t+1}\left(\mathbf{h}_{t+1}^\alpha\right)$, and $\mathbf{m}_{t+1}^\alpha = f\left(\mathbf{a}_{t+1}^\alpha, \mathbf{s}_{t+1}\right)$ for each t.

In words, at each period t, an action a_t is selected according to α_t based on the information history $h_t = (h_{t-1}, a_{t-1}, m_{t-1})$. In turn, its execution generates an outcome m_t that the DM may consider in subsequent periods. Note that α_1 prescribes only one action, $\alpha_1\left(a^0, m^0\right)$, which, together with realization s_1 of $\mathbf{s}_1$, initializes the recursion by sending outcome m_1.

The sequence of σ-algebras $\left(\sigma\left(\mathbf{h}_t^{\alpha}\right)\right)$ on S^{∞} generated by the information process $(\mathbf{h}_t^{\alpha})$ is a filtration that describes the information structure generated and used by strategy α. Since feedback will typically not be perfect, this filtration is coarser than the one generated by the state process $\mathbf{s}^{\infty}$; that is, $\sigma\left(\mathbf{h}_t^{\alpha}\right) \subseteq \sigma\left(\mathbf{s}^{t-1}\right)$ for each $t > 1$. For this reason, without loss of generality, we can regard $\mathbf{h}_t^{\alpha}$ as well as $\mathbf{a}_t^{\alpha}$ and $\mathbf{m}_{t-1}^{\alpha}$ as functions defined on S^{t-1}.[12]

Each finite history $h_t = \left(a^{t-1}, m^{t-1}\right)$ corresponds to the cylinder:

$$I(h_t) = \left\{s^{\infty} \in S^{\infty} : \forall \tau \in \{1, \ldots, t-1\},\, f(a_\tau, s_\tau) = m_\tau\right\} \in \sigma\left(\mathbf{s}^{t-1}\right).$$

This is the information about the realized sequence of states revealed by h_t.

Since states are not directly observed, we can focus on processes $(\mathbf{a}_t^{\alpha})$, $(\mathbf{m}_t^{\alpha})$, and $(\mathbf{h}_t^{\alpha})$, keeping the underlying parameterized probability space $\left(S^{\infty}, \mathcal{B}\left(S^{\infty}\right), p_\theta\right)$ in the background. We write events in terms of the processes observable by the DM. In particular,

$$\left[\mathbf{h}_{t+1}^{\alpha} = \left(a^t, m^t\right)\right] = \begin{cases} I\left(a^t, m^t\right), & \text{if } \forall \tau \in \{1, ..., t\}, \alpha_\tau\left(a^{\tau-1}, m^{\tau-1}\right) = a_\tau, \\ \emptyset, & \text{otherwise.} \end{cases}$$

Example 3 *(Act I).* Assume that only bets on either black or yellow are possible, not on green. As a result, we now have $A = \{b, y\}$ and the table in the Prelude becomes:

f	B	Y	G
b	1	0	0
y	0	1	0

Throughout we will consider two strategies, denoted by α^{NE} (No Experimentation) and α^{E} (Experimentation). Strategy α^{NE} dictates betting on black forever. Strategy α^{E} dictates experimenting with yellow in period 1, and, from period 2 onwards, the action prescribed is constant but depends on the result of this experimentation: If a success is observed in period 1, y is chosen; otherwise b is chosen every period thereafter. Formally:

Strategy α^{NE}: For each $h_t = (a^{t-1}, m^{t-1})$,

$$\alpha_t^{NE}(h_t) = \begin{cases} b \text{ if } t = 1, \\ y \text{ if } t > 1, \text{ and } (y, 1) \in \{(a_1, m_1), \ldots, (a_{t-1}, m_{t-1})\}, \\ b \text{ if } t > 1, \text{ and } (y, 1) \notin \{(a_1, m_1), \ldots, (a_{t-1}, m_{t-1})\}. \end{cases}$$

(Of course, to assess deviations, the strategy must specify actions to be taken at histories that the strategy itself excludes, such as what to do after having bet on yellow.)

By always betting on black, the DM cannot observe the relative frequencies of Y and G. In particular, for each period t and state history s^t,

$$\mathbf{a}_t^{\alpha^{NE}}\left(s^{t-1}\right) = b,$$

$$\mathbf{m}_t^{\alpha^{NE}}\left(s^t\right) = \begin{cases} 1 & \text{if } s_t = B, \\ 0 & \text{if } s_t \in \{Y, G\}, \end{cases}$$

$$\mathbf{h}_{t+1}^{\alpha^{NE}}\left(s^t\right) = \begin{cases} \left(\mathbf{h}_t^{\alpha^{NE}}\left(s^{t-1}\right), b, 1\right) & \text{if } s_t = B, \\ \left(\mathbf{h}_t^{\alpha^{NE}}\left(s^{t-1}\right), b, 0\right) & \text{if } s_t \in \{Y, G\}. \end{cases}$$

[12] Recall that $\sigma\left(\mathbf{s}^0\right)$ is the trivial σ-algebra.

Strategy α^E: *For each* $h_t = (a^{t-1}, m^{t-1})$,

$$\alpha_t^E(h_t) = \begin{cases} y & \text{if } t = 1, \\ y & \text{if } t > 1, \text{ and } (y,1) \in \{(a_1, m_1), \ldots, (a_{t-1}, m_{t-1})\}, \\ b & \text{if } t > 1, \text{ and } (y,1) \notin \{(a_1, m_1), \ldots, (a_{t-1}, m_{t-1})\}. \end{cases}$$

The only difference between this strategy and α^{NE} is the action chosen in the first period. Next we describe the induced processes of actions and outcomes:

$$\mathbf{a}_1^{\alpha^E} = y,$$

$$\mathbf{m}_1^{\alpha^E}(s_1) = \begin{cases} 1 & \text{if } s_1 = Y, \\ 0 & \text{if } s_1 \in \{B, G\}, \end{cases}$$

$$\mathbf{h}_2^{\alpha^E}(s_1) = \begin{cases} (y,1) & \text{if } s_1 = Y, \\ (y,0) & \text{if } s_1 \in \{B, G\}, \end{cases}$$

and, for each $t > 1$ and s^t,

$$\mathbf{a}_t^{\alpha^E}\left(s^{t-1}\right) = \begin{cases} y & \text{if } s_1 = Y, \\ b & \text{else}, \end{cases}$$

$$\mathbf{m}_t^{\alpha^E}\left(s^t\right) = \begin{cases} 1 & \text{if } s_1 = Y \text{ and } s_t = Y, \\ 1 & \text{if } s_1 \in \{B, G\} \text{ and } s_t = B, \\ 0 & \text{else}, \end{cases}$$

$$\mathbf{h}_{t+1}^{\alpha^E}\left(s^t\right) = \begin{cases} \left(\mathbf{h}_t^{\alpha^E}\left(s^{t-1}\right), y, 1\right) & \text{if } s_1 = Y \text{ and } s_t = Y, \\ \left(\mathbf{h}_t^{\alpha^E}\left(s^{t-1}\right), b, 1\right) & \text{if } s_1 \in \{B, G\} \text{ and } s_t = B, \\ \left(\mathbf{h}_t^{\alpha^E}\left(s^{t-1}\right), y, 0\right) & \text{if } s_1 = Y \text{ and } s_t \in \{B, G\}, \\ \left(\mathbf{h}_t^{\alpha^E}\left(s^{t-1}\right), b, 0\right) & \text{if } s_1 \in \{B, G\} \text{ and } s_t \in \{Y, G\}. \end{cases}$$ ▲

3. Models and learning

3.1. Distributions and updating

Prior, predictive, and posterior probabilities A probability measure $\mu : \mathcal{B}(\Theta) \to [0,1]$ is called a prior. A prior induces a predictive distribution $p_\mu \in \Delta(S^\infty)$ defined by $p_\mu(B) = \int_\Theta p_\theta(B)\, \mu(\mathrm{d}\theta)$ for all $B \in \mathcal{B}(S^\infty)$. We make the following maintained assumption about the prior of the DM.

Assumption 2. The prior μ has finite support and satisfies the following "no-surprise" property:

$$\forall t \in \mathbb{N}, \forall h_t \in H_t \quad p_{\bar\theta}(I(h_t)) > 0 \Rightarrow p_\mu(I(h_t)) > 0. \tag{3}$$

In words, $p_{\bar\theta}$-almost surely, the DM is not surprised by what he observes.[13] A sufficient condition for (3) is to have $\mu\left(\bar\theta\right) > 0$. A weaker sufficient condition for (3), one which allows for

[13] Our results would still hold without the "no-surprise" assumption if we endow the DM with a conditional probability system describing how he updates, or revises, his beliefs over models (see the working paper version of Battigalli et al., 2019).

misspecification, is the existence of a model $\hat{\theta} \in \operatorname{supp} \mu$ such that $\bar{\theta}$ is absolutely continuous with respect to $\hat{\theta}$.[14] For example, it is sufficient that the prior assigns positive probability to a full support model. With this, the posterior of the DM after every information history h_t with positive objective probability (i.e., with $p_{\bar{\theta}}(I(h_t)) > 0$) is given by Bayes rule:

$$\forall \theta \in \Theta \quad \mu(\theta | h_t) = \frac{p_\theta(I(h_t))\,\mu(\theta)}{p_\mu(I(h_t))}.$$

Our use of Bayes rule for an ambiguity sensitive DM may seem surprising since Bayesian updating has a revealed-preference foundation in the Subjective Expected Utility axioms for preferences over strategies. However, if we do not allow for commitment, we cannot rely on a revealed-preference approach to justify Bayesian updating in this setting: Strategies cannot be chosen, only actions can be chosen. This impossibility of commitment to a particular strategy is critical in a context where preferences are allowed to be dynamically inconsistent.

By sticking to Bayesian updating, we can preserve the separation between ambiguity attitudes and the perception of ambiguity of the static KMM decision criterion in this dynamic setting. This separation is lost if we consider dynamically consistent rules for updating beliefs.[15] Our approach complements the analysis of the dynamic choices of a KMM decision maker by Hanany et al. (2019). In their work, they maintain dynamic consistency of the preferences of the DM but consider a different updating rule for beliefs.

Observationally equivalent models Given a strategy α and a probability distribution p_θ, let $p_\theta^\alpha : \sigma(\mathbf{h}^\alpha) \to [0,1]$ denote the restriction of p_θ to the σ-algebra generated by the α-observable events: $\sigma(\mathbf{h}^\alpha) = \sigma\left(\cup_t \sigma\left(\mathbf{h}_t^\alpha\right)\right)$. With this, we define the $\sigma\left(\mathbf{h}_t^\alpha\right)$-measurable correspondence (random set) representing the collection of models that are deemed possible and that, conditional on $\mathbf{h}_t^\alpha\left(s^{t-1}\right)$, are *observationally equivalent* to the true model $\bar{\theta}$ under α and prior μ. Formally,[16]

$$\boldsymbol{\Theta}_t^{\alpha,\mu}\left(s^\infty\right) = \left\{\theta \in \operatorname{supp}\mu\left(\cdot|\mathbf{h}_t^\alpha\left(s^{t-1}\right)\right) : p_\theta^\alpha\left(\cdot|\mathbf{h}_t^\alpha\left(s^{t-1}\right)\right) = p_{\bar{\theta}}^\alpha\left(\cdot|\mathbf{h}_t^\alpha\left(s^{t-1}\right)\right)\right\}.$$

Note that, for some s^∞, the set $\boldsymbol{\Theta}_t^{\alpha,\mu}(s^\infty)$ may be empty if $\bar{\theta} \notin \operatorname{supp}\mu$.

The next lemma establishes a monotonicity property of this correspondence. We introduce the following abuse of notation/terminology: When a property holds p_θ-almost surely, we will simply say that it holds θ-almost surely (θ-a.s. for short).

Lemma 1. *For every true model $\bar{\theta} \in \Theta$ and every period t, $\boldsymbol{\Theta}_t^{\alpha,\mu} \subseteq \boldsymbol{\Theta}_{t+1}^{\alpha,\mu}$ $\bar{\theta}$-a.s.*

The intuition behind the lemma is as follows. The set $\boldsymbol{\Theta}_t^{\alpha,\mu}$ may contain models that disagree with $\bar{\theta}$ on the relative probabilities of past events (up to $t-1$), but that agree with $\bar{\theta}$ on the relative probabilities of future events (from t). Almost surely, every model that agrees with $\bar{\theta}$ on future events conditional on information up to $t-1$ also agrees on future events conditional on information up to t. Act II of our running Example shows that the inclusion can be strict.

It follows from the lemma that, $\bar{\theta}$-a.s.,

$$\boldsymbol{\Theta}_1^{\alpha,\mu} := \left\{\theta \in \operatorname{supp}\mu : p_\theta^\alpha = p_{\bar{\theta}}^\alpha\right\} \subseteq \boldsymbol{\Theta}_t^{\alpha,\mu}$$

[14] That is, $\operatorname{supp}\bar{\theta} \subseteq \operatorname{supp}\hat{\theta}$. Under perfect feedback, this condition is equivalent to (3).

[15] See Hanany and Klibanoff (2009), further discussed in Section 7, and Maccheroni et al. (2006).

[16] It is actually enough to require $p_\theta^\alpha\left(E \mid \mathbf{h}_t^\alpha\left(s^{t-1}\right)\right) = p_{\bar{\theta}}^\alpha\left(E \mid \mathbf{h}_t^\alpha\left(s^{t-1}\right)\right)$ for all $E \in \cup_{t\geq 1}\sigma\left(\mathbf{h}_t^\alpha\right)$. That is, observational equivalence is determined by the α-observable events.

for every t. Set $\mathbf{\Theta}_1^{\alpha,\mu}$ represents the irreducible model uncertainty that, when $\bar{\theta}$ is the true model, the DM faces if he plays α and holds belief μ.[17] When $\mathbf{\Theta}_1^{\alpha,\mu} = \operatorname{supp}\mu$, such uncertainty and strategy do not allow any learning, as all the models that the DM initially deems possible are α-observationally equivalent to the true model. The opposite is true when $\mathbf{\Theta}_1^{\alpha,\mu} = \{\bar{\theta}\}$, since in this case the DM will assign probability arbitrarily close to 1 to the true model as he accumulates observations.

In what follows, we will often study properties of a triple $(\alpha, \mu, \bar{\theta})$ where α is the strategy carried out by the DM, μ is his prior belief over models at period 0, and $\bar{\theta}$ is the correct model. We want to understand the behavior and learning of a DM who follows strategy α when his prior is μ and the true model is $\bar{\theta}$. Therefore, we have a particular interest in statements that hold $\bar{\theta}$-a.s., that is, that are almost surely true for the correct model. The notion of observationally equivalent models motivates the following definition.

Definition 1. A triple $(\alpha, \mu, \bar{\theta})$ is *consistent* at time t if $\mathbf{\Theta}_t^{\alpha,\mu} \neq \emptyset$ $\bar{\theta}$-a.s.

In words, a triple $(\alpha, \mu, \bar{\theta})$ is consistent at time t if, conditional on the available information $\mathbf{h}_t^\alpha$, at least one model deemed possible is α-observationally equivalent to the true model.[18]

Let:

$$\sigma_{\geq t}(\mathbf{h}^\alpha(s^{t-1})) = \sigma\left(\cup_{\tau\geq 1}\left\{B \in \sigma\left(\mathbf{h}_{t+\tau}^\alpha\right) : B \subseteq I\left(\mathbf{h}_t^\alpha(s^{t-1})\right)\right\}\right)$$

denote the sigma-algebra of α-observable events from date t onwards given s^{t-1}. Then:

$$\mathbf{\Theta}_t^{\alpha,\mu}\left(s^\infty\right) = \left\{\theta \in \operatorname{supp}\mu\left(\cdot|\mathbf{h}_t^\alpha\left(s^{t-1}\right)\right) : \forall E \in \sigma_{\geq t}\left(\mathbf{h}^\alpha\left(s^{t-1}\right)\right),\ p_\theta^\alpha(E) = p_{\bar{\theta}}^\alpha(E)\right\}.$$

Hence, $(\alpha, \mu, \bar{\theta})$ is consistent at t if, for $\bar{\theta}$-almost every s^{t-1}, there exists some $\theta \in \operatorname{supp}\mu\left(\cdot \mid \mathbf{h}_t^\alpha\left(s^{t-1}\right)\right)$ such that $p_\theta^\alpha(E) = p_{\bar{\theta}}^\alpha(E)$ for all $E \in \sigma_{\geq t}\left(\mathbf{h}^\alpha\left(s^{t-1}\right)\right)$. Of course, an obvious sufficient condition for consistency is that $\mu(\bar{\theta}) > 0$.

In view of Lemma 1, for a triple $(\alpha, \mu, \bar{\theta})$ it is easier to meet the condition for consistency as t gets larger. Let $T(\alpha, \mu, \bar{\theta})$ denote the infimum of the set of t at which $(\alpha, \mu, \bar{\theta})$ is consistent.[19] If $T(\alpha, \mu, \bar{\theta}) = T < \infty$, we say that the triple is *consistent from period* T; if the triple is consistent for some period $T' < \infty$, we say that it is *eventually consistent*. We begin by showing that, under our consistency assumption, beliefs converge almost surely.

Lemma 2. *If* $(\alpha, \mu, \bar{\theta})$ *is eventually consistent, then the process* $\left(\mu\left(\cdot|\mathbf{h}_t^\alpha\right)\right)$ *converges* $\bar{\theta}$*-a.s.*

For the next result, note that convergence of beliefs along a path does not imply convergence of actions. That is, more than one action can be played infinitely often along a path.

[17] In this work, we use the term "belief" to denote the probability assessment over (stochastic) models. Using the terminology of Marinacci (2015), this belief represents how the DM addresses epistemic uncertainty, whereas models capture the (perceived) physical uncertainty.

[18] The word "consistent" may remind the reader of the consistency criterion imposed in Arrow and Green (1973). However, theirs is an "existence of equilibrium condition" requiring that, given any DM's action and true model, there exists a subjective model conceivable by the DM that is observationally equivalent to the actual one.

[19] We set $T(\alpha, \mu, \bar{\theta}) = \infty$ if no such t exists.

Proposition 1. *If* $\left(\alpha, \mu, \bar{\theta}\right)$ *is eventually consistent, then*

$$\lim_{t\to\infty} \mu(\left\{\theta \in \Theta : F\left(\bar{a}, \theta\right) = F\left(\bar{a}, \bar{\theta}\right)\right\} \mid \mathbf{h}_t^{\alpha}\left(s^{\infty}\right)) = 1 \tag{4}$$

for $\bar{\theta}$*-almost every* s^{∞} *and every action* $\bar{a}$ *played infinitely often by* α *on path* s^{∞}.

In words, a triple $(\alpha, \mu, \bar{\theta})$ that is eventually consistent allows the DM to learn the α-observable implications of the true model $\bar{\theta}$ in the long run.[20]

Under perfect feedback, the true model is asymptotically identified. This is the classical result of Doob (1949).

Corollary 1. *If* $\left(\alpha, \mu, \bar{\theta}\right)$ *is eventually consistent, then, under perfect feedback,*

$$\mu\left(\bar{\theta} | \mathbf{h}_t^{\alpha}\right) \to 1 \quad \bar{\theta}\text{-a.s.}$$

Example 4 *(Act II).* Suppose that the DM:

1. knows that $1/3$ of the balls are black (and so all his models θ are such that $\theta\left(B\right) = 1/3$);
2. has a 3-point prior μ with $\operatorname{supp}\mu = \left\{\theta^{Y}, \theta^{uni}, \theta^{G}\right\}$ and believes it is equally likely that the true model is either θ^{Y} (with $\theta^{Y}\left(Y\right) = 2/3$), the uniform model θ^{uni}, or θ^{G} (with $\theta^{G}\left(G\right) = 2/3$):

Marginals	B	Y	G
θ^{Y}	$\frac{1}{3}$	$\frac{2}{3}$	0
θ^{uni}	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
θ^{G}	$\frac{1}{3}$	0	$\frac{2}{3}$

Prior	θ^{Y}	θ^{uni}	θ^{G}
μ	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

By requiring to always bet on the color with known proportion, strategy α^{NE} does not allow the DM to learn anything. Formally,

$$\forall s^{t} \in S^{t}, \qquad \mu\left(\cdot | \mathbf{h}_{t+1}^{\alpha^{NE}}\left(s^{t}\right)\right) = \mu.$$

Here, $T\left(\alpha^{NE}, \mu, \bar{\theta}\right) = 1$ and $\mathbf{\Theta}_1^{\alpha^{NE}, \mu} = \operatorname{supp}\mu$ regardless of what $\bar{\theta}$ is; strategy α^{NE} only allows partial identification. We also have that $T\left(\alpha^{E}, \mu, \bar{\theta}\right) = 1$ for α^{E} if $\bar{\theta} \in \operatorname{supp}\mu$, although α^{E} may allow identification of the true model. After a success in period 1, we have $\operatorname{supp}\mu\left(\cdot | (y, 1)\right) = \{\theta^{Y}, \theta^{\text{uni}}\}$ and:

$$\mathbf{\Theta}_2^{\alpha^{E}, \mu}\left(s^{\infty}\right) = \begin{cases} \left\{\bar{\theta}\right\} & \text{if } \mathbf{h}_2^{\alpha^{E}}\left(s_1\right) = (y, 1), \\ \left\{\theta^{Y}, \theta^{uni}, \theta^{G}\right\} & \text{if } \mathbf{h}_2^{\alpha^{E}}\left(s_1\right) = (y, 0), \end{cases}$$

[20] In this respect we differ from the literature on learning with a misspecified prior, and in particular from Esponda and Pouzo (2016). The standard result of this literature is that, starting from a misspecified prior, beliefs about the model and its observable implications may converge to a distribution of observables that is different from the one implied by the model. Instead, although our notion of consistency allows for misspecification, by Proposition 1 beliefs about observables converge to the distribution implied by the model.

Notice that the inclusion in Lemma 1 can be strict. Strategy α^E experiments with y at $t=1$, but reverts to b if a failure is observed, that is, if $s_1 \in \{B, G\}$. Given α^E, model θ^Y is the only one that predicts success at $t=1$ with probability $2/3$, whereas all models predict success with probability $1/3$ (betting on B) from $t=2$ if a failure is observed at $t=1$. Therefore, if $\bar{\theta} = \theta^Y$,

$$\mathbf{\Theta}_1^{\alpha^E,\mu}\left(s^\infty\right) = \left\{\theta^Y\right\} \subset \left\{\theta^Y, \theta^{uni}, \theta^G\right\} = \mathbf{\Theta}_2^{\alpha^E,\mu}\left(s^\infty\right)$$

for $s_1 = B, G$.

By Proposition 1,

$$\mu\left(\cdot|\mathbf{h}_t^{\alpha^E}\right) \to \begin{cases} \delta_{\bar{\theta}} & \text{if } \mathbf{h}_2^{\alpha^E} = (y,1), \\ \mu\left(\cdot \mid (y,0)\right) & \text{if } \mathbf{h}_2^{\alpha^E} = (y,0), \end{cases}$$

where $\delta_{\bar{\theta}}$ denotes the Dirac measure on $\bar{\theta}$. If experimentation succeeds, the true model is asymptotically learned. Otherwise, if $h_2 = (y,0)$, posterior beliefs attain their limit value as early as the second period, and the DM remains in the dark. ▲

4. Value

We posit that, in the absence of model uncertainty, the DM ranks alternative strategies according to the standard Discounted Expected Utility criterion. Let $R : A \times \Theta \to \mathbb{R}$ denote the (objective) expected reward function defined in Section 2.1.

Assumption 1' *(Discounted expected utility on lotteries).* There exists a constant $\delta \in [0,1)$ such that, for every objective probability measure θ, history h_t with $p_\theta\left(I\left(h_t\right)\right) > 0$, and strategies α and β, the DM prefers α to β if and only if:

$$\begin{aligned} &\sum_{\tau=t}^{\infty} \delta^{\tau-t} \sum_{h_\tau \in H_\tau} R\left(\alpha\left(h_\tau\right), \theta\right) p_\theta\left(\left[\mathbf{h}_\tau^{\alpha|h_t} = h_\tau\right] | h_t\right) \\ \geq &\sum_{\tau=t}^{\infty} \delta^{\tau-t} \sum_{h_\tau \in H_\tau} R\left(\beta\left(h_\tau\right), \theta\right) p_\theta\left(\left[\mathbf{h}_\tau^{\beta|h_t} = h_\tau\right] | h_t\right). \end{aligned}$$

Under model uncertainty, we postulate a dynamic version of smooth ambiguity preferences. Let ϕ be the function capturing ambiguity attitudes introduced in Section 2.1. Given prior μ, if history h_t with $p_\mu\left(I\left(h_t\right)\right) > 0$ is observed the DM ranks strategy α according to the present value of the continuation stream of utility certainty equivalents:[21]

$$V\left(\alpha, \mu \mid h_t\right) := \sum_{\tau=t}^{\infty} \delta^{\tau-t} \phi^{-1}\left(\int_\Theta \phi\left(\sum_{h_\tau \in H_\tau} R\left(\alpha\left(h_\tau\right), \theta\right) p_\theta\left(h_\tau | h_t\right)\right) \mu\left(\mathrm{d}\theta \mid h_t\right)\right). \tag{5}$$

This criterion evaluates uncertain one-period outcomes according to the smooth ambiguity model and then aggregates their (utility) certainty equivalents over time through discounting. Therefore, (utility) smoothing over time is irrelevant. Indeed, when the DM evaluates two continuation streams of utility certainty equivalents, he is interested only in their discounted sum, not on their

[21] We abbreviate $\left(\left[\mathbf{h}_\tau^{\alpha|h_t} = h_\tau\right] | h_t\right)$ as $\left(h_\tau | h_t\right)$.

variability over time. Note that the value only depends on the continuation strategy induced by α starting from h_t.

In particular, we obtain:

(i) $V(\alpha, \mu \mid h_t) = \sum_{\tau=t}^{\infty} \delta^{\tau-t} \left(\sum_{h_\tau \in H_\tau} R\left(\alpha(h_\tau), \sum \mu(\theta|h_t)\,\theta\right) \int_\Theta p_\theta(h_\tau \mid h_t)\,\mu(\mathrm{d}\theta \mid h_t) \right)$ (up to an affine transformation) when ϕ is linear;

(ii) $V(\alpha, \mu \mid h_t) = \sum_{\tau=t}^{\infty} \delta^{\tau-t} \sum_{h_\tau \in H_\tau} R(\alpha(h_\tau), \theta)\, p_\theta(h_\tau|h_t)$ when $\operatorname{supp}\mu = \{\theta\}$.

Our analysis relies on the continuity of V with respect to the prior μ.

Lemma 3. *For every strategy α and history h_t, the functional $V(\alpha, \cdot \mid h_t)$ is continuous on the set of priors μ such that $p_\mu(I(h_t)) > 0$.*

With this lemma, we obtain the following additional corollary to Proposition 1.

Corollary 2. *If $\left(\alpha, \mu, \bar{\theta}\right)$ is eventually consistent, then, $\overline{\theta}$-a.s.,*

$$\left| V\left(\alpha, \mu \mid \mathbf{h}_t^\alpha\right) - V\left(\alpha, \delta_{\bar{\theta}} \mid \mathbf{h}_t^\alpha\right) \right| \to 0.$$

This corollary tells us that, in a consistent triple, the strategy becomes unambiguous on path in terms of value. Of course, this result does not imply that $\mu\left(\cdot \mid \mathbf{h}_t^\alpha\right) \to \delta_{\bar{\theta}}$, only that the present value of the strategy that is used converges to the true value. In particular, even in the limit, alternative strategies may entail unknown outcome distributions. Therefore, this corollary is the dynamic version of the observation made in the static analysis of BCMM that only equilibrium strategies have to be unambiguous.

Note that, except for the benchmark case of ambiguity neutrality, this time-additive value *does not admit a recursive formulation*. This is related to the well-known dynamic inconsistency of decision makers with non-neutral attitudes toward ambiguity. For this reason, we are precluded from employing many of the standard dynamic programming results. We provide an example of these inconsistencies in our setting.

Example 5 *(Dynamic inconsistency).* Consider a modified version of our running example. There are only two periods. Only bets on either black or yellow are possible, not on green. However, by paying a small cost ε, it is also possible to bet on black and to observe the color of the selected ball (action bo). Thus, the outcome has two components: $m = (m_1, m_2)$, where m_1 is the monetary payoff and m_2 is the color of the drawn ball under action bo, and null information (denoted by $*$) under actions b and y. Finally, we normalize payoffs as $u(a, (m_1, m_2)) = m_1$. The feedback function f is thus described by the following table:

f	B	Y	G
b	$1, *$	$0, *$	$0, *$
y	$0, *$	$1, *$	$0, *$
bo	$1-\varepsilon, B$	$0-\varepsilon, Y$	$0-\varepsilon, G$

Suppose that the DM:

1. knows that $1/3$ of the balls are black (and so all her models θ are such that $\theta(B) = 1/3$);
2. believes it is equally likely that the true model is either $\hat{\theta}^Y$ or $\hat{\theta}^G$:

Marginals	B	Y	G
$\hat{\theta}^Y$	$\frac{1}{3}$	$\frac{5}{12}$	$\frac{1}{4}$
$\hat{\theta}^G$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{5}{12}$

Prior	$\hat{\theta}^Y$	$\hat{\theta}^G$
μ	$\frac{1}{2}$	$\frac{1}{2}$

Let $\phi(u) = -e^{-10u}$. Then, the ex-ante optimal strategy if ε is sufficiently small is:

Strategy β: "Bet on black observing the color at $t = 1$. For $t = 2$, given yellow in the first period, bet on yellow, otherwise bet on black."[22] The ex-ante value of strategy β is[23]:

$$V(\beta, \mu) = 0.\overline{3} - \varepsilon + \delta 0.3364.$$

However, β does not satisfy the one-deviation property, i.e., it is not incentive compatible. Indeed, after having observed yellow, the DM prefers to bet on black. The posterior belief after having chosen *bo* and having observed yellow is:

Posterior	$\hat{\theta}^Y$	$\hat{\theta}^G$
$\mu(\cdot\|(bo, Y))$	$\frac{5}{8}$	$\frac{3}{8}$

Hence,

$$V(\beta, \mu \mid (bo, Y)) \cong 0.3207 < \frac{1}{3},$$

where $1/3$ is the value of betting on black.

This is a typical example of dynamically-inconsistent preferences. At period 0, for sufficiently small ε, the DM would want to commit to conditioning his behavior on the observed draw. In particular, he would like to choose y if the draw in the first period is Y, that is, after history (bo, Y). Indeed, even if betting on yellow leads to ambiguous consequences, the DM is confident that with high probability, if $\hat{\theta}^G$ is the true model, Y will not be the first-period draw. Therefore, even under model $\hat{\theta}^G$, this strategy presents a moderately high expected value. However, after having observed (bo, Y), even if the posterior probability of $\hat{\theta}^G$ is lower, the DM considers the consequences of choosing action y to be too ambiguous. Indeed, the expected value under model $\hat{\theta}^G$, $1/4$, is quite small. Therefore, since the DM is highly ambiguity averse, he will select b.

Moreover, it can be shown that the strategy "always bet on black" has a lower ex-ante value, $(1 + \delta)/3$, but satisfies the one-deviation property. A sophisticated DM will not pay the cost ε anticipating that he will not condition his behavior on the observed outcome, even if this conditioning is ex-ante optimal. ▲

4.1. Stationary strategies

A strategy α is *stationary* if, given the prior μ, it depends on history only through the induced posterior belief; that is, for all $t, t' \in \mathbb{N}$ and for every two histories $h_t \in H_t$, $h'_{t'} \in H_{t'}$ such that $p_\mu(I(h_t))$ and $p_\mu(I(h'_{t'}))$ are strictly positive:

[22] Note that this is not a proper strategy since it does not assign an action to every information history. In particular, it does not assign an action to personal histories ruled out by the strategy itself. However, the specification of the actions selected at those information histories is irrelevant in determining ex-ante optimality.

[23] We refer to the working paper version for the computations used to obtain formulas in the examples.

$$\mu(\cdot|h_t) = \mu(\cdot|h'_{t'}) \Rightarrow \alpha(h_t) = \alpha(h'_{t'}).$$

Note that stationarity is a property of the pair (α, μ) of strategy and prior.

The following result, obtained using a coupling argument, shows that when the DM uses a stationary strategy, the function V depends on the history only through belief updating.

Lemma 4. *For all α and μ satisfying stationarity, if h_t and $h'_{t'}$ are two histories such that $p_\mu(I(h_t))$ and $p_\mu\left(I\left(h'_{t'}\right)\right)$ are strictly positive, then $\mu(\cdot \mid h_t) = \mu(\cdot \mid h'_{t'})$ implies $V(\alpha, \mu|h_t) = V\left(\alpha, \mu|h'_{t'}\right)$.*

Under ambiguity neutrality, Hinderer (1970)[24] proves the existence of an optimal stationary strategy for arbitrary beliefs. Therefore, it is without loss of generality to focus on stationary strategies. We cannot adopt this approach "as is" because it relies on a notion of global optimality that may violate incentive compatibility under dynamically inconsistent preferences, as shown in Example 5. However, Proposition 2 in the next section provides a partially analogous result for the notion of rationality (intra-personal equilibrium) used in this paper. Given this result, we focus on stationary strategies in the sequel. Therefore, with an abuse of notation, we will often regard α as a function of beliefs over probability models.[25] More precisely, given prior μ, we call *belief-range of* μ the set of beliefs that the DM may hold with positive probability under p_μ:

$$\left\{\mu(\cdot|h_t) : t \in \mathbb{N}, h_t \in H_t, p_\mu(I(h_t)) > 0\right\}.$$

Then for all $\nu = \mu(\cdot|h_t)$ in the belief range of μ, $\alpha(\nu)$ is defined as:

$$\alpha(\nu) = \alpha(h_t).$$

5. Self-confirming equilibrium and learning

5.1. Self-confirming equilibrium

A self-confirming equilibrium (SCE) is a state in which the DM chooses a subjectively optimal action according to the static smooth criterion and, given this choice, his posterior belief coincides with his prior belief. Formally:

Definition 2. A triple $(a^*, \mu^*, \bar{\theta}) \in A \times \Delta(\Theta) \times \Theta$ is a *self-confirming equilibrium* (SCE) if:

(i) $\mu^*\left(\left\{\theta \in \Theta : F(a^*, \theta) = F\left(a^*, \bar{\theta}\right)\right\}\right) = 1$;
(ii) $a^* \in \arg\max_{a \in A} \bar{V}(a, \mu^*)$.

We say that a^* is an SCE action if it is part of an SCE $(a^*, \mu^*, \bar{\theta})$, and that a^* is a Nash equilibrium action if $a^* \in \arg\max_{a \in A} R\left(a, \bar{\theta}\right)$.

The second condition says that a^* is a (myopic, or one-period) best response to μ^* given the ambiguity attitude determined by ϕ. The first condition is a self-confirming property adapted to

[24] Building on the results of Dynkin (1965).

[25] We refer the interested reader to the working paper version for more general results for non-stationary strategies in a non i.i.d. setting.

the static framework. It is equivalent to requiring that, $\bar{\theta}$-a.s., $\mu^*(\cdot|a^*, f(a^*, \mathbf{s})) = \mu^*(\cdot)$, because all models in the support of μ^* yield the same distribution of outcomes as the true model $\bar{\theta}$ given action a^*. We can interpret this condition as follows. The distribution of outcomes that the DM "observes" in the long run if he always plays a^* is precisely what he expects it to be. In this sense, a^* is *unambiguous* for μ^*; since payoffs are observable, the self-confirming property implies that the expected distribution of payoffs coincides with the one implied by the true model $\bar{\theta}$.

Remark 1. If $(a^*, \mu^*, \bar{\theta})$ is an SCE, $R(a^*, \theta) = \mathbb{E}_{F(a^*,\theta)}[u_{a^*}]$ is constant over suppμ^*. ▲

5.2. Rational learning dynamics

When the DM faces a recurrent choice problem, the notion of SCE characterizes behavior and beliefs after the latter have "converged." In other words, the data provided by the equilibrium strategy do not lead to any further updating because the models that the DM deems possible in an SCE cannot be distinguished from each other or from the true model.

In dynamic settings, we are interested not only in behavior after beliefs have reached a steady state, but also in rational behavior as the DM is learning from the data. To define our notion of rationality, we introduce the following notation. For any information history h_t and action a, let $\alpha/(h_t, a)$ denote the continuation strategy that selects a at h_t and behaves as α thereafter.[26]

Definition 3. A pair (α, μ) is *rational* if it satisfies stationarity and if

$$\forall t \in \mathbb{N}, \forall h_t \in H_t, \forall a \in A \quad p_\mu(I(h_t)) > 0 \Rightarrow V(\alpha, \mu \mid h_t) \geq V(\alpha/(h_t, a), \mu \mid h_t).$$

Besides stationarity, this condition is the one-deviation property, which says that—for every information history h_t that the DM deems reachable with positive probability—action $\alpha(\mu(\cdot \mid h_t))$ maximizes the continuation value conditional on h_t given that α is expected to apply in the future. The motivation is the following: Strategy α is a plan formulated by a sophisticated DM who understands his sequential incentives. In each period t, the DM only controls the action in that period, and therefore we require that he maximizes his value with respect to what he can control, given the predicted behavior of his "future selves," i.e., his continuation strategy. If the time horizon is finite, then this condition is equivalent to folding-back planning: In all periods, the DM predicts to behave in the last period according to his last-period contingent incentives; given such last-period prediction, the DM predicts to behave in the second-to-last period according to his second-to-last period contingent incentives, and so on. When the DM is ambiguity neutral, the one-deviation principle implies that if (α, μ) is rational, then strategy α is subjectively optimal given μ.

A strategy α is said to be *rational given* μ if the pair (α, μ) is rational. The proposition below establishes the existence of rational strategies given any prior.

Proposition 2. *For every prior μ there exists a strategy α that is rational given μ.*

We illustrate the concept of rationality in our running example.

[26] Formally, the value function maps strategies into real numbers. However, we have already pointed out that the value at a history h_t depends only on the continuation of the posited strategy. Therefore, the expression $V(\alpha/(h_t, a), \mu \mid h_t)$ is well defined.

Example 6 *(Act III).* We normalize payoffs as $u(a,0)=0$ and $u(a,1)=1$. Outcomes are thus the bets' payoffs. Moreover, we assume that $\phi(u)=-e^{-\lambda u}$, so that higher (absolute) ambiguity aversion corresponds to higher λ (see Klibanoff et al., 2005).

Suppose that the DM features the prior μ presented in Act II. We consider the same strategies α^{NE} and α^{E} analyzed there. The former strategy involves no experimentation as it recommends always betting on black, the color with the known proportion. Thus, the value of this strategy is independent of histories and beliefs, and it is given by[27]:

$$V(\alpha^{NE},\mu|h_t)=\frac{\frac{1}{3}}{1-\delta}.$$

The latter strategy recommends betting on y at $t=1$ and then switching to b permanently if and only if this first bet is unsuccessful. If the DM chooses y, the outcomes are informative about the distribution, and he updates his belief. Recall by Act II that the DM has a uniform 3-point prior μ with $\operatorname{supp}\mu=\{\theta^{Y},\theta^{uni},\theta^{G}\}$. If we denote $\mu(\cdot|h_t):=\left(\mu\left(\theta^{uni}|h_t\right),\mu\left(\theta^{Y}|h_t\right),\mu\left(\theta^{G}|h_t\right)\right)$, the posterior is:

$$\mu(\cdot|(y,1))=\left(\frac{1}{3},\frac{2}{3},0\right)$$

if the outcome is 1 (success), and:

$$\mu(\cdot|(y,0))=\left(\frac{1}{3},\frac{1}{6},\frac{1}{2}\right)$$

otherwise.

After the first period, strategy α^{E} recommends a fixed action. Thus, the continuation value is equal to $\frac{1}{1-\delta}$ times the single-period expected payoff. Specifically, for any history h_t $(t>1)$ that induces belief $\mu(\cdot|h_t)$, the continuation-value after a success in period 1 is

$$V(\alpha^{E},\mu|h_t)=\frac{\phi^{-1}\left(\mu\left(\theta^{uni}|h_t\right)\phi(\frac{1}{3})+\mu\left(\theta^{Y}|h_t\right)\phi(\frac{2}{3})\right)}{1-\delta}.$$

For the initial (empty) history, we have:

$$V(\alpha^{E},\mu|h_1)=\phi^{-1}\left(\frac{1}{3}\phi(\frac{1}{3})+\frac{1}{3}\phi(\frac{2}{3})+\frac{1}{3}\phi(0)\right)+\frac{\delta}{1-\delta}\phi^{-1}\left(\frac{1}{3}\phi\left(\frac{1}{3}\right)+\frac{1}{3}\phi\left(\frac{5}{9}\right)+\frac{1}{3}\phi\left(\frac{1}{3}\right)\right).$$

Two forces affect the option value of experimentation: ambiguity aversion (the higher the value of λ, the lower the value of experimentation) and patience (the higher the value of δ, the higher the value of experimentation). Given this, strategy α^{NE} is preferred if either $\delta=0$ or λ is high enough given $\delta>0$; if so, the pair $\left(\alpha^{NE},\mu\right)$ is rational. As for strategy α^{E}, if δ is sufficiently high and λ is low enough, e.g., $\lambda=1$ and $\delta=0.39$, strategy α^{E} satisfies the one-deviation property at $\left(a^0,m^0\right)$. However, because of experimentation, we need to consider two different contingencies.

[27] This holds only for histories allowed by the strategy, namely on path.

1. If experimentation is successful (i.e., $s_1 = Y$), the DM learns that model θ^G is false and updates his belief from $(1/3, 1/3, 1/3)$ to $(1/3, 2/3, 0)$. Moreover, at every information history, Bayesian updating implies that the posterior will be of the form $(1-k, k, 0)$, with $k \in (0,1)$. At this point, the strategy recommends sticking to y. It can be checked that this recommendation is better than trying out b once before switching to y thereupon. In other words, it satisfies the one-deviation property: For all $\delta \in (0,1)$ and all $\lambda > 0$, the value at an information history $h_t = ((y,1),...)$ with $\mu(\cdot|h_t) = (1-k, k, 0)$ satisfies:

$$\begin{aligned} V(\alpha^E, \mu|h_t) &= \frac{\phi^{-1}\left((1-k)\,\phi(\theta^{uni}(Y)) + k\phi(\theta^Y(Y))\right)}{1-\delta} \\ &= \frac{\phi^{-1}\left((1-k)\,\phi(\tfrac{1}{3}) + k\phi(\tfrac{2}{3})\right)}{1-\delta} \\ &> \frac{1}{3} + \delta\frac{\phi^{-1}\left((1-k)\,\phi(\tfrac{1}{3}) + k\phi(\tfrac{2}{3})\right)}{1-\delta} \\ &= V(\alpha^E/(h_t, b), \mu|h_t). \end{aligned}$$

2. If experimentation is unsuccessful (i.e., $s_1 \in \{B, G\}$), the posterior lowers the weight of model θ^Y relative to models θ^{uni} and θ^G, so that $p_\mu(Y \mid (y,0)) < p_\mu(B \mid (y,0)) = 1/3$. Thereupon, strategy α^E recommends switching (and sticking) to black, so that the continuation value is the same as that under α^{NE}. Moreover, since betting on black does not lead to any further updating, it is enough to check the inequality with second-period beliefs. For sufficiently small δ, or for sufficiently high λ,

$$V(\alpha^E, \mu|h_2) = \frac{1}{3}\frac{1}{1-\delta} > V\left(\alpha^E/(h_1, y), \mu|h_1\right).$$

In particular, this inequality holds with $\lambda = 1$ and $\delta = 0.39$, and we have already argued that $\left(\alpha^E, \mu\right)$ satisfies the one-deviation property at the initial history; therefore, $\left(\alpha^E, \mu\right)$ is rational. ▲

5.3. Convergence to SCE

We are interested in studying the limit behavior of a rational DM. In particular, we investigate the conditions that imply convergence to SCE. Building on Lemma 2, we provide a learning foundation to the SCE concept of BCMM.

The stochastic process of actions and beliefs $\left(\mathbf{a}_t^\alpha, \mu\left(\cdot|\mathbf{h}_t^\alpha\right)\right)$ converges to an SCE if, for $\bar\theta$-almost every s^∞, beliefs converge to a limit $\mu^\alpha_{s^\infty}$ and there exists a finite time t such that $(\mathbf{a}^\alpha_\tau(s^{\tau-1}), \mu^\alpha_{s^\infty}, \bar\theta)$ forms an SCE for all $\tau \geq t$. Note that the tail sequence of actions $(\mathbf{a}^\alpha_\tau(s^{\tau-1}))_{\tau\geq t}$ is not required to be constant, but each action in the tail is a one-period best reply to the limit belief, which is confirmed given said action.

Proposition 3. *Assume that (α, μ) is rational. If $(\alpha, \mu, \bar\theta)$ is eventually consistent, then the stochastic process of actions and beliefs $\left(\mathbf{a}_t^\alpha, \mu\left(\cdot|\mathbf{h}_t^\alpha\right)\right)$ converges to an SCE.*

The intuition is as follows. Since the action set A is finite, after a certain amount of time, every action chosen by α is played infinitely often. Under the stated assumptions, beliefs converge almost surely to a random limit $\mu^\alpha_{s^\infty}$. Thus, each action chosen by α in the long run must be a myopic best reply to the limit belief. This holds because the updated beliefs converge and the

value of experimentation vanishes for actions played infinitely often. Proposition 1 implies that, for $\bar{\theta}$-almost every s^∞, the limit belief $\mu^\alpha_{s^\infty}$ assigns probability 1 to the models that induce the same probabilities over consequences as $\bar{\theta}$ given the actions played in the long run. Therefore, for every action a^* chosen by α in the long run, $(a^*, \mu^\alpha_{s^\infty}, \bar{\theta})$ must be an SCE.

As noted above, the realized sequence of actions $\left(\mathbf{a}^\alpha_t(s^{t-1})\right)$ need not converge unless there is a unique myopic best reply to the limit belief $\mu^\alpha_{s^\infty}$. If the myopic best reply is a unique action a^*, the action sequence $\left(\mathbf{a}^\alpha_t(s^{t-1})\right)$ is eventually constant at a^* and $\left(a^*, \mu^\alpha_{s^\infty}, \bar{\theta}\right)$ is an SCE. Moreover, after a finite time, the agent chooses an action that maximizes the one-period value given the current belief (and not only limit one); that is, exploration (experimentation) becomes irrelevant, all that matters is exploitation. Formally:

Proposition 4. *Assume that (α, μ) is rational and that $\left(\mathbf{a}^\alpha_t, \mu\left(\cdot|\mathbf{h}^\alpha_t\right)\right)$ converges to an SCE on path s^∞. If:*

$$\arg\max_{a\in A} \phi^{-1}\left(\int_\Theta \phi\left(R\left(a,\theta\right)\right)\mu^\alpha_{s^\infty}(\mathrm{d}\theta)\right) = \left\{a^*\right\}$$

for some $a^* \in A$, there exists some τ such that, for all $t \geq \tau$,

$$\mathbf{a}^\alpha_t(s^{t-1}) = a^* \in \arg\max_{a\in A} \phi^{-1}\left(\int_\Theta \phi\left(R\left(a,\theta\right)\right)\mu\left(\mathrm{d}\theta|\mathbf{h}^\alpha_t\left(s^{t-1}\right)\right)\right).$$

It is important to stress that this convergence implies neither that the limit belief is the Dirac measure supported by the correct model, nor that the limit action is the objective myopic best reply. However, the limit pairs of beliefs and actions almost surely satisfy the standard properties of stochastic limits in the (expected utility) stochastic control literature. Indeed, the realization $\left(a^*, \mu^\alpha_{s^\infty}\right)$ features the following:

- *(Confirmed Beliefs):* $\mu^\alpha_{s^\infty}$ assigns probability 1 to the models that are observationally equivalent to the true $\overline{\theta}$ given a^* (see, Proposition 1);
- *(Subjective Myopic Best Reply):* Even if the discount factor is strictly positive, the agent maximizes his one-period value. That is, exploitation prevails over exploration.

In contrast, in a Nash equilibrium, beliefs are correct (i.e., $\mu = \delta_{\bar{\theta}}$) and the action played is an objective myopic best reply. A sufficient condition for convergence to an SCE where the action played is objectively optimal is to have own-action independence.

Corollary 3. *Under own-action independence, if $(\alpha, \mu, \overline{\theta})$ is eventually consistent and (α, μ) is rational, then the stochastic action process $\left(\mathbf{a}^\alpha_t\right)$ converges to a Nash equilibrium action.*

Note that own-action independence guarantees convergence to a Nash equilibrium under observable payoffs, a maintained assumption in this work. If we relax this hypothesis, the stronger condition of perfect feedback is needed.

Our running example illustrates how the true model may remain unidentified in the limit.

Example 7 *(Act IV).* Consider the strategy α^E of the previous acts. Again, recall that the DM has a uniform 3-point prior μ with $\operatorname{supp}\mu = \left\{\theta^Y, \theta^{uni}, \theta^G\right\}$. In Act II, we show that $(\alpha^E, \mu, \overline{\theta})$

is consistent from period 1, whereas, in Act III, we have proved that with parameters $\lambda = 1$ and $\delta = 0.39$, (α^E, μ) is rational. We can show how our convergence result obtains in this specific case. Suppose that $\overline{\theta} = \theta^Y$. By Lemma 2, beliefs converge. In particular:

$$\mu_{s^\infty}^{\alpha^E} = \left(\mu_{s^\infty}^{\alpha^E}(\theta^{uni}), \mu_{s^\infty}^{\alpha^E}(\theta^Y), \mu_{s^\infty}^{\alpha^E}(\theta^G)\right) = \begin{cases} (\frac{1}{3}, \frac{1}{6}, \frac{1}{2}) & \text{if } s_1 \in \{B, G\}, \\ (0, 1, 0) & \text{if } s_1 = Y. \end{cases}$$

If experimentation is unsuccessful, the posterior of μ lowers the weight of model θ^Y relative to models θ^{uni} and θ^G; thereupon, strategy α^E recommends switching (and sticking) to black, so there is no additional updating. On the other hand, if the experimentation is successful, strategy α^E prescribes sticking to yellow thereupon, and then the correct model θ^Y is asymptotically identified.

If $s_1 \in \{B, G\}$, for every $t > 1$,

$$\mathbf{a}_t^{\alpha^E}(s^{t-1}) = b,$$

and $(b, (1/3, 1/6, 1/2), \theta^Y)$ is the SCE that obtains in the limit. Note that in this case the DM will end up choosing an objectively sub-optimal action.

If $s_1 = Y$,

$$\mathbf{a}_t^{\alpha^E}(s^{t-1}) = y,$$

for every $t > 1$, and $(y, (0, 1, 0), \theta^Y)$ is the SCE that obtains in the limit. It is immediate to see that these actions maximize one-period value for limit beliefs and that the distribution of probabilities over outcomes confirms them.

Finally, consider strategy α^{NE}. In Act III, we argue that $\left(\alpha^{NE}, \mu\right)$ is rational if the DM is sufficiently ambiguity averse. In this case, regardless of the correct model $\bar{\theta} \in \Theta = \left\{\theta^{\,\mathrm{Y}}, \theta^{\,\mathrm{uni}}, \theta^{\,\mathrm{G}}\right\}$, we have almost sure convergence to an SCE from period 1. Indeed, the DM sticks to black from the first period onwards, and black is the myopic best reply to the confirmed prior $\mu = (1/3, 1/3, 1/3)$. However, note that if the correct model is $\theta^{\,\mathrm{Y}}$, betting on black is objectively sub-optimal. ▲

6. Comparative dynamics for changes in ambiguity aversion

6.1. Certainty traps: the general case

Act III of our running example suggests that as ambiguity aversion increases, experimentation becomes less attractive. In this section we formalize this intuition. Actions that induce the same probabilities of payoffs under all the models that the DM deems possible are appealing under ambiguity aversion, but generate no new information about the underlying probability model. To obtain evidence on the correct model, the DM has to choose an action that will potentially induce a different probability measure over payoffs under the different models he deems possible—that is, he has to experiment. An ambiguity averse DM is inclined to avoid such ambiguous actions.

Fix an arbitrary belief ν; we say that action a is ν-*unambiguous* if $F(a, \theta) = F\left(a, \theta'\right)$ for every $\theta, \theta' \in \operatorname{supp} \nu$. In words, an action a is unambiguous given the DM's beliefs if all models entertained by the DM assign the same probabilities to outcomes given a. Otherwise, we say that a is ν-ambiguous. The next proposition shows that, if a strategy α is rational given μ for an ambiguity-neutral DM and prescribes unambiguous actions, then every strategy β that is rational given μ for a strictly ambiguity averse DM must also prescribe unambiguous actions. In other

words, if experimentation is not rational for an ambiguity neutral DM, then it cannot be rational for an ambiguity averse DM with the same beliefs. For notational simplicity, in this section we will assume that the utility function u is injective.

Proposition 5. *Assume that (α, μ) is rational under ambiguity neutrality and (β, μ) is rational under strict ambiguity aversion. For every belief ν in the belief-range of μ, if $\alpha(\nu)$ is ν-unambiguous, then the same holds for $\beta(\nu)$.*

In what follows, we restrict our attention to cases where there is a *unique μ-ambiguous action*.[28] Although this assumption is restrictive, it encompasses interesting stochastic control problems such as two-armed bandits with a safe arm. Moreover, this restriction parallels the one needed for important comparative statics results in the case of choice under risk.[29]

The next proposition establishes that, given a prior μ, the sequence of actions and posteriors induced by a strategy β that is rational under strict ambiguity aversion will almost surely converge faster to an SCE (given the true model) than the sequence of actions and posterior beliefs corresponding to a strategy α that is rational under ambiguity neutrality. The intuition is as follows. Under the stated assumptions, the decision problem amounts to deciding how long to experiment, choosing the unique ambiguous action, say a^*. Assuming that convergence occurs, there are two possibilities: Either the DM never stops experimenting, or he stops at some finite time t, choosing the best unambiguous action thereafter. In the first case, convergence to an SCE occurs (typically) at infinity. In the second case, it occurs at the stopping time. By Proposition 5, if an ambiguity neutral DM prefers to stop at time t, then a strictly ambiguity averse DM prefers to stop as well. Alternatively, if an ambiguity neutral DM chooses a^* forever, an ambiguity averse DM either does the same, or starts choosing the best unambiguous action from some period t onwards and is henceforth "trapped" in a self-confirming equilibrium.

Proposition 6. *Assume that (i) there is a unique μ-ambiguous action a^*, (ii) (α, μ) is rational under ambiguity neutrality and (β, μ) is rational under strict ambiguity aversion, and (iii) $\left(\alpha, \mu, \bar{\theta}\right)$ is eventually consistent. Then, $\bar{\theta}$-a.s., the action-belief process $(\beta(\mu(\cdot|\mathbf{h}_t^\beta)), \mu(\cdot|\mathbf{h}_t^\beta))$ converges to an SCE at least as fast as the action-belief process $\left(\alpha\left(\mu\left(\cdot|\mathbf{h}_t^\alpha\right)\right), \mu\left(\cdot|\mathbf{h}_t^\alpha\right)\right)$.*

The next proposition shows that an ambiguity averse DM is less likely than an ambiguity neutral DM to eventually play the Nash equilibrium (i.e., objectively optimal) action.

Proposition 7. *Assume that (i) there is a unique μ-ambiguous action a^*, (ii) (α, μ) and (β, μ) are, respectively, rational under ambiguity neutrality and strict ambiguity aversion, and (iii) $\left(\alpha, \mu, \bar{\theta}\right)$ and $\left(\beta, \mu, \bar{\theta}\right)$ are eventually consistent. Then, $\bar{\theta}$-a.s., if $\left(\mathbf{a}_t^\beta\right)$ converges to a Nash Equilibrium action, so does $\left(\mathbf{a}_t^\alpha\right)$.*

We illustrate the previous results in the classical setup introduced by Rothschild (1974) where a monopolist trades-off exploration of the demand curve for his good against exploitation using the price that (subjectively)-maximizes one period profit.

[28] We refer to Battigalli et al. (2019) for an in-depth analysis of the role of this assumption.

[29] Specifically, the seminal contribution by Arrow (1971) shows that, when there is a unique risky asset, the amount invested in such an asset is decreasing in the risk aversion of the DM. Additional reasons to focus on this case are provided by Proposition 8.

Example 8. A monopolist who is uncertain about the demand for its product faces a new customer each period. The cost of producing a unit is $c > 0$. The monopolist can charge either a low price, p_L, or a high price, p_H, where $p_H > p_L > c \geq 0$. Each new customer has a reservation price in $\{p_H, p_L, 0\}$, and he buys the product if and only if his reservation price is weakly larger than the ask price. If the price is set to p_i and a sale is made, the monopolist's profit is $p_i - c$; otherwise, the profit is 0. Thus, we are considering build-to-order production. Here, $A = \{p_L, p_H\}$, $M = \{0, 1\}$, $S = \{p_H, p_L, 0\}$,

$$f(a, s) = \begin{cases} 1 & a \leq s, \\ 0 & a > s, \end{cases}$$

and:

Payoff $u(p, m)$	$m = 1$	$m = 0$
$p = p_L$	$p_L - c$	0
$p = p_H$	$p_H - c$	0

Suppose that according to his prior μ, the monopolist believes that $\theta = (\theta(p_H), \theta(p_L), \theta(0))$ is either $\theta_1 := (0.8, 0.1, 0.1)$, or $\theta_2 := (0, 0.9, 0.1)$, and that these two models are equally likely. Here, the monopolist is certain that if he posts the low price he sells with probability 0.9, but he is uncertain about the selling probability at a high price, which is—therefore—a μ-ambiguous action.

Moreover, suppose that the correct model is θ_1 and that $0.9(p_L - c) < 0.8(p_H - c)$. Our previous results imply that an ambiguity averse monopolist will stop experimentation with the high price earlier and that he will be more likely to be trapped in the (objectively) suboptimal SCE where he posts the low price (the μ-unambiguous action). ▲

At first sight, the previous results may seem surprising. Indeed, if a DM is ambiguity averse, why does he not experiment more, so as to eliminate (or reduce) the uncertainty about the true model? This reasoning tacitly relies on a different notion of experimentation. It is true that, typically, an ambiguity averse DM is willing to pay more to eliminate model uncertainty (see e.g., Theorem 2 in Anderson, 2012). However, in our active-learning setting, the DM cannot simply buy information about the true model; learning happens only when actions with ambiguous probabilities of consequences are chosen. Since an ambiguity averse DM dislikes those actions, he will end up resolving less ambiguity than his ambiguity neutral counterpart.

The results above relate to the findings in Anderson (2012). On the theoretical side, his Theorem 1 for two-armed bandits with a safe arm strictly relates to our Proposition 5. The main difference is that Anderson (2012) implicitly assumes the possibility to commit to a strategy. Indeed, the Gittins indices used in that paper characterize the *ex-ante* optimal strategy for a decision maker, that is, the strategy that maximizes the value at the initial history. However, like us, he assumes the DM performs Bayesian updating, a feature that paired with ambiguity aversion induces dynamically inconsistent preferences.

On the experimental side, the theoretical predictions of our model are consistent with the findings presented in Anderson (2012): The behavior of a Subjective Expected Utility maximizer cannot explain joint data about willingness to pay for information about the stochastic process characterizing the ambiguous arm and the amount of experimentation that is performed. In particular, the resulting experimentation is too low, which is the prediction of our model under ambiguity aversion.

Another reason to focus on the case of a unique ambiguous action is to illustrate how our analysis adds to BCMM. Indeed, the following proposition shows that, when there is a unique ambiguous action, the set of SCE actions is *invariant* with respect to the (positive) degree of ambiguity aversion captured by the (concave) function ϕ. Yet, as Propositions 6 and 7 show, ambiguity aversion has "dynamic" effects on the persistence of experimentation and the distribution of long-run outcomes.

Proposition 8. *Assume that there is a unique μ-ambiguous action a^*. For every concave and strictly increasing ϕ, ϕ' and every action $\bar{a}$, if $\left(\bar{a}, \mu, \bar{\theta}\right)$ is an SCE under ambiguity attitudes ϕ, then, for some belief μ', $\left(\bar{a}, \mu', \bar{\theta}\right)$ is an SCE under ambiguity attitudes ϕ'.*

Our running example illustrates. As the unique ambiguous action is to bet on yellow, Proposition 8 implies that the monotonicity result of BCMM (that the SCE set is weakly increasing in the degree of ambiguity aversion) holds vacuously: The set of equilibrium actions is not affected by ambiguity attitudes. Nonetheless, if the true model is θ_Y, the example shows that beliefs converge to the true model with positive probability under ambiguity neutrality, while the process of actions and beliefs is trapped in a non-Nash SCE under ambiguity aversion.

The next example illustrates the relevance of the assumption of a unique ambiguous action.

Example 9 *(Multiple ambiguous actions).* There are four possible states, $S = \left\{g, \bar{g}, b, \bar{b}\right\}$, and two possible models in Θ, the good model θ_g and the bad model θ_b defined as follows:

$$\theta_g\left(g\right) = 0.9 = \theta_b\left(b\right) \text{ and } \theta_g\left(\bar{g}\right) = 0.1 = \theta_b\left(\bar{b}\right).$$

The DM has three actions: He can bet aggressively (action a), bet conservatively (action c), or not bet at all (action n). The feedback received by the DM is his monetary payoff. We also assume risk-neutrality: For all $\bar{a} \in A$, $m \in M$, $u\left(a, m\right) = m$. Feedback and payoffs are summarized in the following table:

f	g	$\bar{g}$	b	$\bar{b}$
a	10	0	0	10
c	5	5	4	4
n	4.2	4.2	4.2	4.2

Therefore,

$$R\left(a, \theta_g\right) = 9,\ R\left(a, \theta_b\right) = 1,\ R\left(c, \theta_g\right) = 5 \text{ and } R\left(c, \theta_b\right) = 4.$$

For simplicity, suppose $\delta = 0$ and $\mu\left(\theta_g\right) = \mu\left(\theta_b\right) = 1/2$. In this case, in the first period, an ambiguity neutral DM bets aggressively (a):

$$\frac{R\left(a, \theta_g\right) + R\left(a, \theta_b\right)}{2} = 5 > 4.5 = \frac{R\left(c, \theta_g\right) + R\left(c, \theta_b\right)}{2}.$$

Similarly, one can check that a DM with intermediate ambiguity attitudes (i.e., $\lambda = 1$) and with the same belief μ bets conservatively. Now, suppose that the true model is θ_g. First note that, since c perfectly reveals the model, the ambiguity averse DM discovers the true model at the end of the first period and starts to bet aggressively from the second period. In other words, convergence to Nash equilibrium (since a is the objectively optimal action under θ_g) happens in one period and with probability 1. Now, consider the ambiguity neutral DM. He bets on a in the

first period, and with probability $0.1 = \theta_g\,(\bar{g})$ he receives message $m = 0$ and updates his beliefs to:

$$\mu\left(\theta_g \mid (a, 0)\right) = 0.1.$$

With this, from the second period onwards, he stops betting (i.e., he chooses n) and remains in the dark. Thus, there is at least probability 0.1 that the ambiguity neutral DM will be trapped in a non-Nash SCE. ▲

In this example there is a misalignment between the most informative action (c) and the most ambiguous one (a). Therefore, by avoiding the most ambiguous action, an ambiguity averse DM quickly gathers information about the true model. We believe that this kind of misalignment is unlikely to arise in applications. Still, we conjecture that the results of this section can be extended to the case of multiple ambiguous actions by first giving an adequate definition of an ambiguity order (see, e.g., Jewitt and Mukerji, 2017), and then imposing a condition of comonotonicity between informativeness and ambiguity of actions.[30]

6.2. Certainty traps: myopic decision makers

Sharper versions of our comparative dynamics results can be provided when the DM is myopic, i.e., when $\delta = 0$. This follows from the fact that, in the present framework where each period consists of a one-stage decision problem, a myopic DM is not vulnerable to dynamic inconsistencies. Despite this simplification, the behavior over time of a myopic DM evolves in interesting ways as he gathers information about the true stochastic process. Indeed, several models of learning in games use the assumption of myopic players (see, e.g., Fudenberg and Kreps, 1995, and Fudenberg and Levine, 1998). We show that when the DM is myopic our comparative statics results hold for the entire spectrum of ambiguity attitudes.

For the rest of this section, we assume that the value is given by:

$$V(\alpha, \mu \mid h_t) := \phi^{-1}\left(\int_{\Theta} \phi\left(R\left(\alpha\left(h_\tau\right), \theta\right)\right) \mu\left(\mathrm{d}\theta \mid h_t\right)\right).$$

We let ϕ' be a strictly increasing and concave transformation of ϕ, i.e., we assume that the DM with ambiguity attitudes ϕ' is *strictly more ambiguity averse* than the one with ambiguity attitudes ϕ.

Proposition 9. *Let (α, μ) be rational under ambiguity attitudes ϕ and let (β, μ) be rational under ambiguity attitudes ϕ'. Then, for every belief ν in the belief-range of μ, if $\alpha(\nu)$ is ν-unambiguous, the same holds for $\beta(\nu)$.*

Similarly, the speed of convergence to the SCE is monotone for the entire spectrum of ambiguity attitudes.

Proposition 10. *Assume that there is a unique μ-ambiguous action a^*; let (α, μ) be rational under ambiguity attitudes ϕ and let (β, μ) be rational under ambiguity attitudes ϕ'; furthermore, assume that $\left(\alpha, \mu, \bar{\theta}\right)$ is eventually consistent. Then, $\bar{\theta}$-a.s., the action-belief process*

[30] We can prove this for the case of a myopic DM ($\delta = 0$).

$\left(\beta\left(\mu\left(\cdot|\mathbf{h}_t^{\beta}\right)\right), \mu\left(\cdot|\mathbf{h}_t^{\beta}\right)\right)$ *converges to an SCE at least as fast as the action-belief process* $\left(\alpha\left(\mu\left(\cdot|\mathbf{h}_t^{\alpha}\right)\right), \mu\left(\cdot|\mathbf{h}_t^{\alpha}\right)\right)$.

As a consequence, the probability of converging to a Nash equilibrium is decreasing in ambiguity aversion as well.

Proposition 11. *Assume that there is a unique* μ*-ambiguous action* a^**. Let* (α, μ) *and* (β, μ) *be rational under ambiguity attitudes* ϕ *and* ϕ'*, respectively, and let* $\left(\alpha, \mu, \bar{\theta}\right)$ *and* $\left(\beta, \mu, \bar{\theta}\right)$ *be eventually consistent. Then,* $\bar{\theta}$*-a.s., if* $\left(\mathbf{a}_t^{\beta}\right)$ *converges to a Nash Equilibrium action, so does* $\left(\mathbf{a}_t^{\alpha}\right)$.

7. Concluding remarks

The concept of self-confirming equilibrium with standard expected utility maximizing agents has been given a rigorous learning foundation. We note that the literature on stochastic control problems implicitly addresses this issue, showing that the behavior and beliefs of an ambiguity neutral agent, who faces an unknown i.i.d. process of states affecting the outcome of his actions, almost surely converges to what we call an SCE.[31] As for games against other agents, convergence cannot be taken for granted; but if it occurs, the limit point must be an SCE (e.g., Fudenberg and Levine, 1993, and Fudenberg and Kreps, 1995).

This learning foundation cannot be mechanically applied to the case of non-neutral ambiguity attitudes. Ambiguity averse agents typically have dynamically inconsistent preferences over strategies, and dynamic inconsistency prevents us from applying standard dynamic programming techniques. Given such difficulties, to derive results and insights about convergence to SCE under ambiguity aversion, we focus on the case of repeated play against nature, assuming that the decision maker is sophisticated and thus takes future incentives into account as he chooses actions in earlier periods. With this, we obtain a result of convergence to SCE under ambiguity aversion (Proposition 3).

We point out that, in several interesting problems, the set of SCE actions is independent of ambiguity attitudes (Proposition 8). Yet, ambiguity aversion affects the dynamics: Higher ambiguity aversion tends to decrease experimentation and therefore makes convergence to Nash equilibrium (best reply to the correct model) less likely. In particular, we show that ambiguity aversion may make it more likely that the agent falls into a suboptimal "certainty trap" whereby he keeps choosing an unambiguous action from which he cannot learn, which leads him to settle faster and prevents him from identifying the objectively optimal action (Propositions 5 through 11).

The adoption of the smooth ambiguity model allows us to separate ambiguity attitudes, a personal feature of the DM, from the perception of ambiguity, which depends on his beliefs about statistical models and therefore changes as new evidence accumulates. We model the process of updating such beliefs in a standard Bayesian fashion, as we regard the chain rule, hence Bayesian updating, as part or rational cognition (cf. Section 3 and Battigalli et al., 2019). Yet, the decision-theoretic literature does not take a clear stand on whether ambiguity averse players should update beliefs according to the standard rules of conditional probabilities (see, for example, Epstein and Schneider, 2007; Hanany and Klibanoff, 2009). We remark that one may conduct an analysis

[31] See Easley and Kiefer (1988) and Section 5 of the working paper version of this article for a detailed analysis of the connection with their work.

similar to ours by considering a DM who uses the Hanany and Klibanoff (2009) updating rule for beliefs. In particular, we can prove that every SCE is stable according to their updating rule. Therefore, the main message of Section 6 that "ambiguity aversion makes convergence to a non-Nash equilibrium more likely" still holds.[32]

We can give a game-theoretic interpretation of our analysis within a population-game scenario. In this setting, the DM recognizes to be unable to influence the actions of future co-players. Nevertheless, experimentation is valuable for him, since a better understanding of the correct distribution of behaviors in co-players' populations may allow him to select a better strategy in the following periods. Fudenberg and Levine (1993) put forward a model of this kind. The main difference with their work is that they consider an overlapping generations model with finitely lived agents. Since we assume an infinite horizon, we have to slightly modify our model by introducing a constant probability of death to embed our analysis in an overlapping generations model, as in Fudenberg and He (2018).[33]

Acknowledgments

We thank Federico Bobbio, Roberto Corrao, Carlo Cusumano, Nicodemo De Vito, Ignacio Esponda, Francesco Fabbri, Drew Fudenberg, Filippo Massari, David Ruiz Gomez, Ran Spiegler, Muhamet Yildiz, and anonymous referees for useful comments and suggestions. Special thanks go to Simone Cerreia-Vioglio and Fabio Maccheroni with whom this project started. Financial support from the European Research Council (advanced grant 324219), Guido Cazzavillan Scholarship, and the AXA Research Fund is gratefully acknowledged.

Appendix A. Proofs and related material

For our proofs, it is often convenient to use the notation $R(a,\theta)=\sum_{s\in S} r(a,s)\,\theta(s)$, where $r: A\times S\to\mathbb{R}$ is the payoff (or reward) function $r(a,s):=(u_a\circ f_a)(s)$.

Lemma 5. *If we endow $\Delta(\Theta)$ with the topology of weak convergence of measures then, for every $a\in A$, the functional $\bar{V}(a,\cdot)$ is continuous.*

Proof. Note that $R(a,\cdot)$ is a bounded function, since $|R(a,\cdot)|\le \max_{s\in S}|r(a,s)|$. Moreover, it is an affine function on a finite dimensional space, so it is continuous. Thus, $\int_\Theta \phi(R(a,\theta))(\cdot)(\mathrm{d}\theta)$ is continuous. Since ϕ is strictly increasing and continuous on the interval

$$\left[\min_{s\in S} r(a,s), \max_{s\in S} r(a,s)\right],$$

ϕ^{-1} is continuous as well. Since $\bar{V}(a,\cdot)=\phi^{-1}\circ\int_\Theta \phi(R(a,\theta))(\cdot)(\mathrm{d}\theta)$, the result follows. □

A.1. Models and learning

Instrumental for the following proofs is the correspondence $\iota_t^\alpha: S^{t-1}\to\sigma(\mathbf{h}_t^\alpha)$ defined by:

$$\iota_t^\alpha\left(s^{t-1}\right):=I\left(\mathbf{h}_t^\alpha\left(s^{t-1}\right)\right)\times S^\infty=\left\{\bar{s}^\infty\in S^\infty:\mathbf{h}_t^\alpha\left(\bar{s}^{t-1}\right)=\mathbf{h}_t^\alpha\left(s^{t-1}\right)\right\}. \tag{6}$$

[32] A formal statement and proof are available by request.

[33] See also Blanchard (1985).

We can regard ι_t^α as the identification correspondence determined by α at time t. This correspondence models the information about state histories which is available ex-ante at time t to a DM who is acting according to strategy α. Clearly, $s^{t-1} \times S^\infty \in \iota_t^\alpha\left(s^{t-1}\right)$, and so the correspondence induces a partition of S^{t-1}. We have *perfect* (*state*) *identification* under α when $\iota_t^\alpha\left(s^{t-1}\right) = \left\{s^{t-1}\right\} \times S^\infty$ for each s^{t-1} and each $t > 1$; in this case, the DM knows the actual past history s^{t-1}. Otherwise, we have partial identification. This dependence on α of the identification correspondence plays a key role in our results. Of course, there is no such dependence under own-action independence of feedback, in which case we can write $\iota_t\left(s^{t-1}\right)$; in particular, under perfect feedback, $\iota_t\left(s^{t-1}\right) = \left\{s^{t-1}\right\} \times S^\infty$.

Proof of Lemma 1. Fix s^t with $p_{\bar\theta}\left(s^t\right) > 0$. Note that $s^t \in \iota_{t+1}^\alpha\left(s^t\right) \subseteq \iota_t^\alpha\left(s^{t-1}\right)$; thus, $p_{\bar\theta}\left(\iota_t^\alpha\left(s^{t-1}\right)\right) \geq p_{\bar\theta}\left(\iota_{t+1}^\alpha\left(s^t\right)\right) \geq p_{\bar\theta}\left(s^t\right) > 0$. Let $\theta \in \mathbf{\Theta}_t^{\alpha,\mu}(s^\infty)$; By definition $\theta \in \operatorname{supp}\mu\left(\cdot \mid \mathbf{h}_t^\alpha\left(s^{t-1}\right)\right)$, and so $p_\theta\left(\iota_t^\alpha\left(s^{t-1}\right)\right) > 0$. We want to show that $\theta \in \mathbf{\Theta}_{t+1}^{\alpha,\mu}(s^\infty)$. To this end, notice that if $E \cap \iota_{t+1}^\alpha\left(s^t\right) = \varnothing$, $p_\theta^\alpha\left(E \mid \mathbf{h}_{t+1}^\alpha\left(s^t\right)\right) = p_{\bar\theta}^\alpha\left(E \mid \mathbf{h}_{t+1}^\alpha\left(s^t\right)\right) = 0$. Therefore, it is enough to show that $p_\theta^\alpha\left(\cdot \mid \mathbf{h}_{t+1}^\alpha\left(s^t\right)\right) = p_{\bar\theta}^\alpha\left(\cdot \mid \mathbf{h}_{t+1}^\alpha\left(s^t\right)\right)$ for $E \in \sigma\left(\mathbf{h}^\alpha\right)$ with $E \subseteq \iota_{t+1}^\alpha\left(s^t\right)$. Fix such an E. Since $\theta \in \mathbf{\Theta}_t^{\alpha,\mu}(s^\infty)$, then:

$$\frac{p_\theta(E)}{p_\theta\left(\iota_t^\alpha\left(s^{t-1}\right)\right)} = \frac{p_{\bar\theta}(E)}{p_{\bar\theta}\left(\iota_t^\alpha\left(s^{t-1}\right)\right)}; \quad \frac{p_\theta\left(\iota_{t+1}^\alpha\left(s^t\right)\right)}{p_\theta\left(\iota_t^\alpha\left(s^{t-1}\right)\right)} = \frac{p_{\bar\theta}\left(\iota_{t+1}^\alpha\left(s^t\right)\right)}{p_{\bar\theta}\left(\iota_t^\alpha\left(s^{t-1}\right)\right)}.$$

The second equality implies $p_\theta\left(\iota_{t+1}^\alpha\left(s^t\right)\right) > 0$. Since:

$$\frac{p_\theta(E)}{p_\theta\left(\iota_{t+1}^\alpha(s^t)\right)} \frac{p_\theta\left(\iota_{t+1}^\alpha(s^t)\right)}{p_\theta\left(\iota_t^\alpha(s^{t-1})\right)} = \frac{p_{\bar\theta}(E)}{p_{\bar\theta}\left(\iota_{t+1}^\alpha(s^t)\right)} \frac{p_{\bar\theta}\left(\iota_{t+1}^\alpha(s^t)\right)}{p_{\bar\theta}\left(\iota_t^\alpha(s^{t-1})\right)},$$

it follows that:

$$p_\theta^\alpha\left(E \mid \mathbf{h}_{t+1}^\alpha\left(s^t\right)\right) = \frac{p_\theta(E)}{p_\theta\left(\iota_{t+1}^\alpha(s^t)\right)} = \frac{p_{\bar\theta}(E)}{p_{\bar\theta}\left(\iota_{t+1}^\alpha(s^t)\right)} = p_{\bar\theta}^\alpha\left(E \mid \mathbf{h}_{t+1}^\alpha\left(s^t\right)\right).$$

Hence, $p_\theta^\alpha\left(\cdot \mid \mathbf{h}_{t+1}^\alpha\left(s^t\right)\right) = p_{\bar\theta}^\alpha\left(\cdot \mid \mathbf{h}_{t+1}^\alpha\left(s^t\right)\right)$. Since $\theta \in \operatorname{supp}\mu\left(\cdot \mid \mathbf{h}_t^\alpha\left(s^{t-1}\right)\right)$ and $p_\theta\left(\iota_{t+1}^\alpha\left(s^t\right)\right) > 0$, it follows that $\theta \in \operatorname{supp}\mu\left(\cdot \mid \mathbf{h}_{t+1}^\alpha\left(s^t\right)\right)$, hence $\theta \in \mathbf{\Theta}_{t+1}^{\alpha,\mu}(s^\infty)$. □

Lemma 6. *For every $\hat\theta \in \Theta$, the process $\left(\mu\left(\hat\theta|\mathbf{h}_t^\alpha\right)\right)$ is a uniformly bounded martingale in $\left(S^\infty, \mathcal{B}\left(S^\infty\right), p_\mu\right)$.*

Proof. Uniform boundedness is immediate from the fact that the process is $[0, 1]$-valued. For every t, we want to show that, for all $k \in \mathbb{R}$ such that $p_\mu\left(\left[\mu\left(\hat\theta|\mathbf{h}_{t-1}^\alpha\right) = k\right]\right) > 0$ we have:

$$\mathbb{E}_{p_\mu}\left[\mu\left(\hat\theta|\mathbf{h}_t^\alpha\right) \middle| \mu\left(\hat\theta|\mathbf{h}_{t-1}^\alpha\right) = k\right] = k.$$

Let h_{t-1} be an arbitrary element of H_{t-1} such that $\mu\left(\hat\theta|h_{t-1}\right) = k$ and $p_\mu\left(h_{t-1}\right) > 0$. By the Law of Iterated Expectations, it is enough to prove that:

$$\mathbb{E}_{p_\mu}\left[\mu\left(\hat\theta|\mathbf{h}_t^\alpha\right) \middle| \mathbf{h}_{t-1}^\alpha = h_{t-1}\right] = k.$$

Recall that the Bayes map yields:

$$\Delta(\Theta) \times A \times M \to \Delta(\Theta)$$
$$(\mu, a, m) \mapsto B(\mu, a, m)(\theta) = \frac{F(a,\theta)(m)\,\mu(\theta)}{\sum_{\theta' \in \operatorname{supp}\mu} F(a,\theta')(m)\,\mu(\theta')}$$

for each m deemed possible according to μ given action a, that is, each m such that the denominator is positive. Define:

$$M(h_{t-1}) = \left\{ m \in M : \sum_{\theta' \in \operatorname{supp}\mu(\cdot|h_{t-1})} F\left(a,\theta'\right)(m)\,\mu\left(\theta'|h_{t-1}\right) > 0 \right\}.$$

With this,

$$\begin{aligned}
&\mathbb{E}_{p_\mu}\left[\mu\left(\hat{\theta}\,|\mathbf{h}_t^\alpha\right)|\mathbf{h}_{t-1}^\alpha = h_{t-1}\right] \\
&= \sum_{m \in M(h_{t-1})} p_\mu\left[\mathbf{m}_{t-1}^\alpha = m | \mathbf{h}_{t-1}^\alpha = h_{t-1}\right] B\left(\mu(\cdot|h_{t-1}), \alpha(h_{t-1}), m\right)\left(\hat{\theta}\right) \\
&= \sum_{m \in M(h_{t-1})} \sum_{\theta \in \operatorname{supp}\mu} \frac{\mu(\theta)\, p_\theta\left(\left\{s^\infty \in I(h_{t-1}) : s_t \in f^{-1}_{\alpha(h_{t-1})}(m)\right\}\right)}{\sum_{\theta' \in \operatorname{supp}\mu} \mu(\theta') p_{\theta'}(I(h_{t-1}))} \\
&\quad \times B\left(\mu(\cdot|h_{t-1}), \alpha(h_{t-1}), m\right)\left(\hat{\theta}\right) \\
&= \sum_{m \in M(h_{t-1})} \sum_{\theta \in \operatorname{supp}\mu} \frac{\mu(\theta)\, p_\theta(I(h_{t-1}))\, F(\alpha(h_{t-1}),\theta)(m)}{\sum_{\theta' \in \operatorname{supp}\mu} \mu(\theta') p_{\theta'}(I(h_{t-1}))} \\
&\quad \times B\left(\mu(\cdot|h_{t-1}), \alpha(h_{t-1}), m\right)\left(\hat{\theta}\right) \\
&= \sum_{m \in M(h_{t-1})} \sum_{\theta \in \operatorname{supp}\mu} F(\alpha(h_{t-1}),\theta)(m)\,\mu(\theta|h_{t-1}) \\
&\quad \times \frac{k F\left(\alpha(h_{t-1}), \hat{\theta}\right)(m)}{\sum_{\theta' \operatorname{supp}\mu(\cdot|h_{t-1})} \mu(\theta'|h_{t-1})\, F(\alpha(h_{t-1}),\theta')(m)} \\
&= \sum_{m \in M(h_{t-1})} k F\left(\alpha(h_{t-1}), \hat{\theta}\right)(m) \frac{\sum_{\theta \in \operatorname{supp}\mu} \mu(\theta|h_{t-1})\, F(\alpha(h_{t-1}),\theta)(m)}{\sum_{\theta' \operatorname{supp}\mu(\cdot|h_{t-1})} \mu(\theta'|h_{t-1})\, F(\alpha(h_{t-1}),\theta')(m)} \\
&= \sum_{m \in M(h_{t-1})} k F\left(\alpha(h_{t-1}), \hat{\theta}\right)(m) = k,
\end{aligned}$$

where the first equality comes from the definition of expected value, the second by the definition of p_μ, the third from the fact that the environment is i.i.d., the fourth from the definition of Bayes map, and the fifth from rearranging the terms.

For the last equality notice that either $\mu\left(\hat{\theta}|h_{t-1}\right) = k = 0$, and the equality is trivial, or $\mu\left(\hat{\theta}|h_{t-1}\right) > 0$, and therefore $m \notin M(h_{t-1})$ implies that $F\left(\alpha(h_{t-1}), \hat{\theta}\right)(m) = 0$. □

Lemma 7. *For every $\hat{\theta} \in \operatorname{supp}\mu$, the process $\left(\mu\left(\cdot|\mathbf{h}_t^\alpha\right)\right)_{t\in\mathbb{N}_0}$ converges $\hat{\theta}$-a.s. to a random limit $\mu^\alpha_{\mathbf{s}^\infty}$.*

Proof. By Lemma 6, the stochastic process $\left(\mu\left(\cdot|\mathbf{h}_t^{\alpha}\right)\right)_{t\in\mathbb{N}_0}$ is a uniformly-bounded martingale in $\left(S^{\infty},\mathcal{B}\left(S^{\infty}\right),p_{\mu}\right)$. By the Martingale Convergence Theorem (Billingsley, 2012, Theorem 35.5), the limit random variable $\mu_{\mathbf{s}^{\infty}}^{\alpha}$ exists p_{μ}-almost surely. This means that there exists a set $E\in\mathcal{B}(S^{\infty})$ such that $p_{\mu}(E)=1$, so $p_{\mu}(S^{\infty}\backslash E)=0$, and:

$$\lim_{t\to+\infty}\mu\left(\cdot|\mathbf{h}_t^{\alpha}\left(s^{\infty}\right)\right)=\mu_{s^{\infty}}^{\alpha}$$

for every $s^{\infty}\in E$. Note that, since $\mu(\widehat{\theta})>0$,

$$\sum_{\theta\in\operatorname{supp}\mu}p_{\theta}(S^{\infty}\backslash E)\mu(\theta)=p_{\mu}(S^{\infty}\backslash E)=0$$

implies that $p_{\widehat{\theta}}(S^{\infty}\backslash E)=0$ and so $p_{\widehat{\theta}}(E)=1$. □

Proof of Lemma 2. Let $T\left(\alpha,\mu,\bar{\theta}\right)=T$. If $\overline{\theta}\in\operatorname{supp}\mu$ the result follows immediately from Lemma 7. If $\overline{\theta}\notin\operatorname{supp}\mu$, by consistency the set $O:=\left\{s^{\infty}:\mathbf{\Theta}_T^{\alpha,\mu}(s^{\infty})\neq\emptyset\right\}$ has $p_{\bar{\theta}}$-probability 1. Define the set $E^{*}:=\left\{s^{\infty}:\exists\nu\in\Delta\left(\Theta\right):\lim_{t\to+\infty}\mu\left(\cdot|\mathbf{h}_t^{\alpha}\left(s^{\infty}\right)\right)=\nu\right\}$. For every $s^{\infty}\in O$, we can find some $\theta(s^{\infty})\in\operatorname{supp}\mu\left(\cdot|\mathbf{h}_T^{\alpha}\left(s^{T-1}\right)\right)$ such that $p_{\overline{\theta}}(A)=p_{\theta(s^{\infty})}(A)$ for every $A\in\sigma_{\geq T}\left(\mathbf{h}_T^{\alpha}\left(s^{T-1}\right)\right)$. Repeating the arguments of Lemmata 6 and 7 with the probability space $\left(I\left(\mathbf{h}_T^{\alpha}\left(s^{T-1}\right)\right),\sigma_{\geq T}\left(\mathbf{h}_T^{\alpha}\left(s^{T-1}\right)\right),\mu\left(\cdot|\mathbf{h}_T^{\alpha}\left(s^{T-1}\right)\right)\right)$ in place of $\left(S^{\infty},\mathcal{B}\left(S^{\infty}\right),\mu\right)$, and $p_{\theta(s^{\infty})}(\cdot|\mathbf{h}_T^{\alpha}\left(s^{T-1}\right))$ in place of $p_{\hat{\theta}}$, we obtain that there exists a set $E\in\sigma_{\geq T}\left(\mathbf{h}_T^{\alpha}\left(s^{T-1}\right)\right)$ such that $p_{\theta(s^{\infty})}(E|\mathbf{h}_T^{\alpha}\left(s^{T-1}\right))=1$, and:

$$\forall s^{\infty}\in E,\exists\nu\in\Delta\left(\Theta\right):\lim_{t\to+\infty}\mu\left(\cdot|\mathbf{h}_t^{\alpha}\left(s^{\infty}\right)\right)=\nu.$$

Moreover,

$$p_{\overline{\theta}}(E^{*}|\mathbf{h}_T^{\alpha}\left(s^{T-1}\right))\geq p_{\overline{\theta}}(E|\mathbf{h}_T^{\alpha}\left(s^{T-1}\right))=p_{\theta(s^{\infty})}(E|\mathbf{h}_T^{\alpha}\left(s^{T-1}\right))=1,$$

and therefore $p_{\overline{\theta}}(E^{*}|\mathbf{h}_T^{\alpha}\left(s^{T-1}\right))=1$. Since s^{∞} was an arbitrary element of the set O with $p_{\overline{\theta}}\left(O\right)=1$, by the Law of Iterated Expectations (see, e.g., 9.7i in Williams, 1991) $p_{\overline{\theta}}(E^{*})=1$. □

For every path s^{∞}, denote by $\mathbf{a}_{\infty}^{\alpha}\left(s^{\infty}\right)$ the set of actions played infinitely often (i.o.) under strategy α along this path. Also for every $\theta\in\Theta$ let $E_{\hat{\theta}}^{\alpha}$ denote the event that an action and a state that occur i.o. also occur jointly i.o. More formally,

$$E_{\hat{\theta}}^{\alpha}=\left\{s^{\infty}:\forall\left(a,\bar{s}\right)\in\mathbf{a}_{\infty}^{\alpha}\left(s^{\infty}\right)\times\operatorname{supp}\hat{\theta},\left(\alpha\left(\mathbf{h}_{t-1}^{\alpha}\left(s^{\infty}\right)\right),s_t\right)=\left(a,\bar{s}\right)\text{ i.o.}\right\}.$$

Lemma 8. *For every $\hat{\theta}\in\operatorname{supp}\mu$, we have $p_{\hat{\theta}}(E_{\hat{\theta}}^{\alpha})=1$.*

Proof. For every $(a,\bar{s})\in\mathbf{a}_{\infty}^{\alpha}\left(s^{\infty}\right)\times\operatorname{supp}\hat{\theta}$, denote by $E(a,\overline{s},n)\subseteq S^{\infty}$ the set of sequences s^{∞} such that $a\in\mathbf{a}_{\infty}^{\alpha}\left(s^{\infty}\right)$ but $\left(\alpha\left(\mathbf{h}_{t-1}^{\alpha}\left(s^{\infty}\right)\right),s_t\right)\neq\left(a,\bar{s}\right)$ for every $t\geq n$. We have:

$$S^{\infty}\backslash E_{\hat{\theta}}^{\alpha}\subseteq\bigcup_{a\in\mathbf{a}_{\infty}^{\alpha}(s^{\infty})}\bigcup_{\bar{s}\in\operatorname{supp}\hat{\theta}}\bigcup_{n\in\mathbb{N}}E(a,\overline{s},n).$$

In turn, for every $k\in\mathbb{N}$,

$$E(a,\overline{s},n)\subseteq\bigcap_{j=1}^{k}\left\{s^{\infty}:\exists s_j\in S\backslash\{\overline{s}\},t_j\geq n,\left(\alpha\left(\mathbf{h}_{t_j-1}^{\alpha}\left(s^{\infty}\right)\right),s_{t_j}\right)=\left(a,s_j\right)\right\},$$

and so:

$$\begin{aligned}p_{\widehat{\theta}}(E(a,\overline{s},n))&\leq\times_{j=1}^{k}p_{\widehat{\theta}}\left(\left\{s^{\infty}:\exists s_j\in S\backslash\{\overline{s}\},t_j\geq n,\left(\alpha\left(\mathbf{h}_{t_j-1}^{\alpha}\left(s^{\infty}\right)\right),s_{t_j}\right)=\left(a,s_j\right)\right\}\right)\\&\leq\times_{j=1}^{k}\widehat{\theta}\left(S\backslash\{\overline{s}\}\right)\\&=\left(\widehat{\theta}\left(S\backslash\{\overline{s}\}\right)\right)^{k}.\end{aligned}$$

Since $\bar{s}\in\operatorname{supp}\hat{\theta}$, this inequality implies that $p_{\widehat{\theta}}(E(a,\overline{s},n))=0$. It follows that $p_{\widehat{\theta}}(S^{\infty}\backslash E_{\hat{\theta}}^{\alpha})=0$, or $p_{\widehat{\theta}}(E_{\hat{\theta}}^{\alpha})=1$. □

Lemma 9. *For every $\hat{\theta}\in\operatorname{supp}\mu$, we have, $\hat{\theta}$-a.s.,*

$$\forall a\in\mathbf{a}_{\infty}^{\alpha}\left(s^{\infty}\right),\ \mu_{s^{\infty}}^{\alpha}\left(\left\{\theta\in\Theta:F\left(a,\theta\right)=F\left(a,\hat{\theta}\right)\right\}\right)=1.$$

Proof. Fix any $\hat{\theta}\in\operatorname{supp}\mu$. Let E be as in the proof of Lemma 7. Define the set $\overline{E}_{\widehat{\theta}}^{\alpha}:=E\cap E_{\widehat{\theta}}^{\alpha}$, and fix a sample path $s^{\infty}\in\overline{E}_{\widehat{\theta}}^{\alpha}$. Suppose by way of contradiction that there is some $a\in\mathbf{a}_{\infty}^{\alpha}\left(s^{\infty}\right)$ and some $m\in M$ such that, for some $\theta\in\operatorname{supp}\mu_{s^{\infty}}^{\alpha}$, we have $F\left(a,\theta\right)(m)\neq F\left(a,\hat{\theta}\right)(m)$. This implies that, for any $\overline{s}\in f_a^{-1}(m)$, we have $B\left(\mu_{s^{\infty}}^{\alpha},a,f(a,\overline{s})\right)\neq\mu_{s^{\infty}}^{\alpha}$; thus, we can find some $\varepsilon>0$ such that $\left\|B\left(\mu_{s^{\infty}}^{\alpha},a,f(a,\overline{s})\right)-\mu_{s^{\infty}}^{\alpha}\right\|=2\varepsilon$. By continuity of the Bayes map at $\left(\mu_{s^{\infty}}^{\alpha},a,f(a,\overline{s})\right)$, there exists some $\delta>0$ such that $\left\|\mu\left(\cdot|\mathbf{h}_{t-1}^{\alpha}\left(s^{\infty}\right)\right)-\mu_{s^{\infty}}^{\alpha}\right\|<\delta$ implies[34]:

$$\left\|B\left(\mu\left(\cdot|\mathbf{h}_{t-1}^{\alpha}\left(s^{\infty}\right)\right),a,f(a,\overline{s})\right)-B\left(\mu_{s^{\infty}}^{\alpha},a,f(a,\overline{s})\right)\right\|<\varepsilon.$$

Therefore,

$$\begin{aligned}2\varepsilon&=\left\|B\left(\mu_{s^{\infty}}^{\alpha},a,f(a,\overline{s})\right)-\mu_{s^{\infty}}^{\alpha}\right\|\\&\leq\left\|B\left(\mu\left(\cdot|\mathbf{h}_{t-1}^{\alpha}\left(s^{\infty}\right)\right),a,f(a,\overline{s})\right)-\mu_{s^{\infty}}^{\alpha}\right\|\\&\quad+\left\|B\left(\mu\left(\cdot|\mathbf{h}_{t-1}^{\alpha}\left(s^{\infty}\right)\right),a,f(a,\overline{s})\right)-B\left(\mu_{s^{\infty}}^{\alpha},a,f(a,\overline{s})\right)\right\|\\&<\left\|B\left(\mu\left(\cdot|\mathbf{h}_{t-1}^{\alpha}\left(s^{\infty}\right)\right),a,f(a,\overline{s})\right)-\mu_{s^{\infty}}^{\alpha}\right\|+\varepsilon,\end{aligned}$$

so $\left\|B\left(\mu\left(\cdot|\mathbf{h}_{t-1}^{\alpha}\left(s^{\infty}\right)\right),a,f(a,\overline{s})\right)-\mu_{s^{\infty}}^{\alpha}\right\|>\varepsilon$. Invoking Lemma 7 and since $s^{\infty}\in\overline{E}_{\widehat{\theta}}$, there exists a sequence of dates $(t_n)_{n\in\mathbb{N}}$ such that, for every n,

$$\left(\alpha\left(\mathbf{h}_{t_n-1}^{\alpha}\left(s^{\infty}\right)\right),s_{t_n}\right)=(a,\bar{s})\ \text{ and }\ \left\|\mu\left(\cdot|\mathbf{h}_{t_n-1}^{\alpha}\left(s^{\infty}\right)\right)-\mu_{s^{\infty}}^{\alpha}\right\|<\delta,$$

and so:

$$\begin{aligned}&\left\|\mu\left(\cdot|\mathbf{h}_{t_n}^{\alpha}\left(s^{\infty}\right)\right)-\mu_{s^{\infty}}^{\alpha}\right\|\\&=\left\|B\left(\mu\left(\cdot|\mathbf{h}_{t_n-1}^{\alpha}\left(s^{\infty}\right)\right),\alpha\left(\mathbf{h}_{t_n-1}^{\alpha}\left(s^{\infty}\right)\right),f\left(\alpha\left(\mathbf{h}_{t_n-1}^{\alpha}\left(s^{\infty}\right)\right),s_{t_n}\right)\right)-\mu_{s^{\infty}}^{\alpha}\right\|\\&=\left\|B\left(\mu\left(\cdot|\mathbf{h}_{t_n-1}^{\alpha}\left(s^{\infty}\right)\right),a,f(a,\overline{s})\right)-\mu_{s^{\infty}}^{\alpha}\right\|>\varepsilon.\end{aligned}$$

[34] Notice that the Bayes map is continuous at all (ν,a,m) such that $\sum_{\theta}F\left(a,\theta\right)(m)\,\nu\left(\theta\right)>0$. Since $s^{\infty}\in\overline{E}_{\widehat{\theta}}^{\alpha}$, this is the case for $\left(\mu_{s^{\infty}}^{\alpha},a,f(a,\overline{s})\right)$.

This contradicts Lemma 7. Since by Lemmata 7 and 8 $p_{\widehat{\theta}}\left(\overline{E}^{\alpha}_{\widehat{\theta}}\right)=1$, the result follows. □

Proof of Proposition 1. The result is obtained from Lemma 9 using the same argument used to derive Lemma 2 from Lemma 7, replacing the set E^* with:

$$E^{*}:=\left\{s^{\infty}:\forall\bar{a}\in\mathbf{a}^{\alpha}_{\infty}\left(s^{\infty}\right),\mu^{\alpha}_{s^{\infty}}\left(\left\{\theta\in\Theta:F\left(\bar{a},\theta\right)=F\left(\bar{a},\bar{\theta}\right)\right\}\right)=1\right\}.$$

Further details are omitted. □

Proof of Corollary 1. Since under perfect feedback

$$\forall\bar{a}\in A\quad\left\{\theta\in\Theta:F\left(\bar{a},\theta\right)=F\left(\bar{a},\bar{\theta}\right)\right\}=\bar{\theta}$$

and A is finite, the statement follows from Proposition 1. □

A.2. Value

In the rest of the Appendix, we will make use of the fact that the dependence on the state in the value can be made explicit. Indeed, by definition we have:

$$\sum_{h_{\tau}\in H_{\tau}}R\left(\alpha\left(h_{\tau}\right),\theta\right)p_{\theta}\left(\left[\mathbf{h}^{\alpha}_{\tau}=h_{\tau}\right]|h_{t}\right)=\sum_{s^{\tau}\in S^{\tau}}r\left(\mathbf{a}^{\alpha}_{\tau}\left(s^{\tau-1}\right),s_{\tau}\right)p_{\theta}(s^{\tau}|h_{t}),$$

so that the value function can be expressed as:

$$V\left(\alpha,\mu\mid h_{t}\right)=\sum_{\tau=t}^{\infty}\delta^{\tau-t}\phi^{-1}\left(\int_{\Theta}\phi\left(\sum_{s^{\tau}\in S^{\tau}}r\left(\mathbf{a}^{\alpha}_{\tau}\left(s^{\tau-1}\right),s_{\tau}\right)p_{\theta}\left(s^{\tau}|h_{t}\right)\right)\mu\left(\mathrm{d}\theta\mid h_{t}\right)\right).$$

Proof of Lemma 3. It is immediate to see that the map:

$$W_{\bar{t}}:\theta\mapsto\phi\left(\sum_{s^{\bar{t}}\in S^{\bar{t}}}r\left(\mathbf{a}^{\alpha}_{\bar{t}}\left(s^{\bar{t}-1}\right),s_{\bar{t}}\right)p_{\theta}\left(s^{\bar{t}}\mid h_{t}\right)\right)$$

is continuous and bounded by $\max_{(a,s)\in A\times S}|\phi\left(r\left(a,s\right)\right)|$. Moreover, as argued in the proof of Lemma 5, ϕ^{-1} is continuous. Since the space of measures endowed with the topology of weak convergence is metrizable, it is enough to establish sequential continuity. By continuity of Bayesian updating with respect to positive probability events and $p_{\mu}\left(I\left(h_{t}\right)\right)>0$,

$$\mu_{n}\to\mu\Rightarrow\mu_{n}\left(\cdot\mid h_{t}\right)\to\mu\left(\cdot\mid h_{t}\right).$$

Therefore, by definition of weak convergence of measures and continuity of $W_{\bar{t}}$ and ϕ^{-1}:

$$\mu_{n}\left(\cdot\mid h_{t}\right)\to\mu\left(\cdot\mid h_{t}\right)\Rightarrow\phi^{-1}\left(\int_{\Theta}W_{\bar{t}}\left(\theta\right)\mu_{n}\left(\mathrm{d}\theta\mid h_{t}\right)\right)\to\phi^{-1}\left(\int_{\Theta}W_{\bar{t}}\left(\theta\right)\mu\left(\mathrm{d}\theta\mid h_{t}\right)\right).$$

Let $\varepsilon>0$. Since $\delta<1$ and W_{τ} is bounded, there exists a T such that for every $\mu\left(\cdot\mid h_{t}\right)$,

$$\left|\sum_{\tau=T}^{\infty}\delta^{\tau-t}\phi^{-1}\left(\int_{\Theta}W_{\tau}\left(\theta\right)\mu\left(\mathrm{d}\theta\mid h_{t}\right)\right)\right|<\varepsilon.$$

But then, let n be such that for every $\tau \leq T$,

$$|\phi^{-1}\left(W_{\tilde{t}}\left(\theta\right)\mu_n\left(\mathrm{d}\theta \mid h_t\right)\right) - \phi^{-1}\left(W_{\tilde{t}}\left(\theta\right)\mu\left(\mathrm{d}\theta \mid h_t\right)\right)| < \varepsilon.$$

It follows that:

$$\begin{aligned}
&|V\left(\alpha, \mu \mid h_t\right) - V\left(\alpha, \mu_n \mid h_t\right)| \\
&= \left|\sum_{\tau=t}^{\infty} \delta^{\tau-t}\phi^{-1}\left(\int_{\Theta} W_\tau\left(\theta\right)\mu\left(\mathrm{d}\theta \mid h_t\right)\right) - \sum_{\tau=t}^{\infty} \delta^{\tau-t}\phi^{-1}\left(\int_{\Theta} W_\tau\left(\theta\right)\mu_n\left(\mathrm{d}\theta \mid h_t\right)\right)\right| \\
&\leq \left|\sum_{\tau=t}^{T} \delta^{\tau-t}\phi^{-1}\left(\int_{\Theta} W_\tau\left(\theta\right)\mu\left(\mathrm{d}\theta \mid h_t\right)\right) - \sum_{\tau=t}^{T} \delta^{\tau-t}\phi^{-1}\left(\int_{\Theta} W_\tau\left(\theta\right)\mu_n\left(\mathrm{d}\theta \mid h_t\right)\right)\right| \\
&\quad + \left|\sum_{\tau=T+1}^{\infty} \delta^{\tau-t}\phi^{-1}\left(\int_{\Theta} W_\tau\left(\theta\right)\mu\left(\mathrm{d}\theta \mid h_t\right)\right)\right. \\
&\quad \left. - \sum_{\tau=T+1}^{\infty} \delta^{\tau-t}\phi^{-1}\left(\int_{\Theta} W_\tau\left(\theta\right)\mu_n\left(\mathrm{d}\theta \mid h_t\right)\right)\right| \\
&< \left(T+2\right)\varepsilon.
\end{aligned}$$

Since ε has been chosen arbitrarily, the result is proved. □

Proof of Corollary 2. By Proposition 1 and Lemma 2, the set:

$$E^* = \left\{s^\infty : \exists \nu \in \Delta\left(\Theta\right) : \lim_{t\to+\infty} \mu\left(\cdot|\mathbf{h}_t^\alpha\left(s^\infty\right)\right) = \nu, \nu(\left\{\theta \in \Theta : F\left(a,\theta\right) = F\left(a,\bar{\theta}\right)\right\}) = 1\right\}$$

has $\overline{\theta}$-probability 1. For each $s^\infty \in E^*$, consider:

$$\begin{aligned}
&\lim_{t\to\infty} V\left(\alpha, \mu \mid \mathbf{h}_t^\alpha\left(s^\infty\right)\right) = \lim_{t\to\infty} \sum_{\tau=t}^{\infty} \delta^{\tau-t}\phi^{-1}\left(\int_{\Theta} W_\tau\left(\theta\right)\mu_{s^\infty}^\alpha\left(\mathrm{d}\theta\right)\right) \\
&= \lim_{t\to\infty} V\left(\alpha, \delta_{\bar{\theta}} \mid \mathbf{h}_t^\alpha\left(s^\infty\right)\right),
\end{aligned}$$

where the first equality follows from Lemma 3, and the second equality follows from the definition of E^*. □

Proof of Lemma 4. First note that since $p_\mu(I\left(h_t\right))$ and $p_\mu(I\left(h'_{t'}\right))$ are strictly positive, then:

$$\frac{\mu\left(\theta\right)p_\theta\left(I\left(h'_{t'}\right)\right)}{p_\mu(I\left(h'_{t'}\right))} = \mu(\theta \mid h'_{t'}) = \mu(\theta \mid h_t) = \frac{\mu\left(\theta\right)p_\theta\left(I\left(h_t\right)\right)}{p_\mu(I\left(h_t\right))}.$$

In particular,

$$\mu(\theta \mid h'_{t'}) = \mu(\theta \mid h_t) > 0 \Rightarrow p_\theta\left(I\left(h_t\right)\right) > 0, \text{ and } p_\theta\left(I\left(h'_{t'}\right)\right) > 0.$$

That is, the models in the support of the $\mu(\cdot \mid h'_{t'}) = \mu(\cdot \mid h_t)$ assign positive probability to the two conditioning events. In turn, this implies that $p_\theta(\cdot|h_t)$ and $p_\theta(\cdot|h_{t'})$ are obtained by Bayes rule. Hence we have:

$$V(\alpha, \mu \mid h_t)$$
$$= \sum_{\tau=t}^{\infty} \delta^{\tau-t} \phi^{-1} \left(\int_{\operatorname{supp} \mu(\cdot|h_t)} \phi \left(\sum_{s^\tau \in S^\tau} r\left(\mathbf{a}_\tau^\alpha\left(s^{\tau-1}\right), s_\tau\right) p_\theta(s^\tau|h_t) \right) \mu(\mathrm{d}\theta \mid h_t) \right).$$

To show our result, we will prove that for every n in $\mathbb{N}_0$,

$$\phi^{-1} \left(\int_{\operatorname{supp} \mu(\cdot|h_t)} \phi \left(\sum_{s^{t+n} \in S^{t+n}} r\left(\mathbf{a}_{t+n}^\alpha\left(s^{t+n-1}\right), s_{t+n}\right) p_\theta(s^{t+n}|h_t) \right) \mu(\mathrm{d}\theta \mid h_t) \right)$$
$$= \phi^{-1} \left(\int_{\operatorname{supp} \mu(\cdot|h'_{t'})} \phi \left(\sum_{s^{t'+n} \in S^{t'+n}} r\left(\mathbf{a}_{t'+n}^\alpha\left(s^{t'+n-1}\right), s_{t'+n}\right) p_\theta(s^{t'+n}|h'_{t'}) \right) \mu\left(\mathrm{d}\theta \mid h'_{t'}\right) \right).$$

Since $V(\alpha, \mu \mid h_t)$ and $V\left(\alpha, \mu \mid h'_{t'}\right)$ are defined as the discounted sum from $n = 0$ to infinity of, respectively, the first and second line above, the statement will follow.

Let $n \in \mathbb{N}_0$, $\theta \in \operatorname{supp} \mu(\cdot \mid h_t) = \operatorname{supp} \mu(\cdot \mid h'_{t'})$, and $(k_0, ..., k_n) \in S^{n+1}$ such that $\theta(k_i) \neq 0$ for every $i \in \{0, ..., n\}$. Define:

$$K(k_0, ..., k_n) := \left\{ s^{t+n} \in I(h_t) \mid s_t = k_0, ..., s_{t+n} = k_n \right\}$$

and

$$K'(k_0, ..., k_n) := \left\{ s^{t'+n} \in I\left(h'_{t'}\right) \mid s_{t'} = k_0, ..., s_{t'+n} = k_n \right\}.$$

By definition of p_θ, for every $s^{t+n} \in K(k_0, ..., k_n)$ and every $s^{t'+n} \in K'(k_0, ..., k_n)$,

$$p_\theta(s^{t+n}|h_t) = \prod_{i=0}^{n} \theta(k_n) = p_\theta(s^{t'+n}|h'_{t'}).$$

To ease notation, fix $(k_0, ..., k_n)$ momentarily and let $K = K(k_0, ..., k_n)$ and $K' = K'(k_0, ..., k_n)$. We show that

$$r\left(\mathbf{a}_{t+n}^\alpha\left(s^{t+n-1}\right), s_{t+n}\right)$$

is constant on K. Indeed, we prove by way of induction that for every $j \in \{0, ..., n\}$, $\mathbf{a}_{t+j}^\alpha\left(s^{t+j-1}\right)$ is constant on K. Since for every $s^{t+n} \in K$ we have $\mathbf{h}_t^\alpha(s^{t-1}) = h_t$,

$$\mathbf{a}_t^\alpha\left(s^{t-1}\right) = \alpha\left(\mu\left(\cdot|h_t\right)\right).$$

Suppose by way of induction that the statement holds for $j' \leq j$. Thus, for every $s^{t+n} \in K$ we have:

$$\mathbf{h}_{t+j}^\alpha(s^{t+j-1}) = \left(h_t, \mathbf{a}_t^\alpha\left(s^{t-1}\right), f\left(\mathbf{a}_t^\alpha\left(s^{t-1}\right), s_t\right), ..., \mathbf{a}_{t+j-1}^\alpha\left(s^{t+j-2}\right),\right.$$
$$\left. f\left(\mathbf{a}_{t+j-1}^\alpha\left(s^{t+j-2}\right), s_{t+j-1}\right)\right),$$

which, by definition of K and by the inductive hypothesis, is constant on K. It follows that:

$$\mathbf{a}_{t+j}^\alpha\left(s^{t+j-1}\right) = \alpha\left(\mu\left(\cdot|\mathbf{h}_{t+j}^\alpha(s^{t+j-1})\right)\right)$$

is constant on K. Since $s_{t+n} = k_n$ for every $s^{t+n} \in K$, we have shown that $r\left(\mathbf{a}^{\alpha}_{t+n}\left(s^{t+n-1}\right), s_{t+n}\right)$ is also constant on K. A similar argument shows that $r\left(\mathbf{a}^{\alpha}_{t'+n}\left(s^{t'+n-1}\right), s_{t'+n}\right)$ is constant on K'. Moreover, for every s^{t+n} in K and $s^{t'+n}$ in K', and for every j in $\{0, ..., n\}$,

$$\mu\left(\cdot|\mathbf{h}^{\alpha}_{t+j}(s^{t+j-1})\right) = \mu\left(\cdot|\mathbf{h}^{\alpha}_{t'+j}(s^{t'+j-1})\right).$$

We prove this equality by induction on j. By hypothesis, it is true for $j = 0$. Let $j \in \{1, ..., n\}$ and suppose that it is true for $j-1$. This implies that:

$$\begin{aligned}\mathbf{a}^{\alpha}_{t+j-1}\left(s^{t+j-2}\right) &= \alpha\left(\mu\left(\cdot|\mathbf{h}^{\alpha}_{t+j-1}(s^{t+j-2})\right)\right)\\ &= \alpha\left(\mu\left(\cdot|\mathbf{h}^{\alpha}_{t'+j-1}(s^{t'+j-2})\right)\right)\\ &= \mathbf{a}^{\alpha}_{t'+j-1}\left(s^{t'+j-2}\right).\end{aligned}$$

Therefore:

$$\begin{aligned}&\mu\left(\theta|\mathbf{h}^{\alpha}_{t+j}(s^{t+j-1})\right)\\ &= \frac{\mu\left(\theta|\mathbf{h}^{\alpha}_{t+j-1}(s^{t+j-2})\right) F\left(\mathbf{a}^{\alpha}_{t+j-1}\left(s^{t+j-2}\right), \theta\right)\left(f\left(\mathbf{a}^{\alpha}_{t+j-1}\left(s^{t+j-2}\right), k_{j-1}\right)\right)}{F\left(\mathbf{a}^{\alpha}_{t+j-1}\left(s^{t+j-2}\right), \theta_{\mu\left(\cdot|\mathbf{h}^{\alpha}_{t+j-1}(s^{t+j-2})\right)}\right)\left(f\left(\mathbf{a}^{\alpha}_{t+j-1}\left(s^{t+j-2}\right), k_{j-1}\right)\right)}\\ &= \frac{\mu\left(\theta|\mathbf{h}^{\alpha}_{t'+j-1}(s^{t'+j-2})\right) F\left(\mathbf{a}^{\alpha}_{t'+j-1}\left(s^{t'+j-2}\right), \theta\right)\left(f\left(\mathbf{a}^{\alpha}_{t'+j-1}\left(s^{t'+j-2}\right), k_{j-1}\right)\right)}{F\left(\mathbf{a}^{\alpha}_{t'+j-1}\left(s^{t'+j-2}\right), \theta_{\mu\left(\cdot|\mathbf{h}^{\alpha}_{t'+j-1}(s^{t'+j-2})\right)}\right)\left(f\left(\mathbf{a}^{\alpha}_{t'+j-1}\left(s^{t'+j-2}\right), k_{j-1}\right)\right)}\\ &= \mu\left(\theta|\mathbf{h}^{\alpha}_{t'+j-1}(s^{t'+j-1})\right).\end{aligned}$$

This in turn implies that, for every s^{t+n} in K and $s^{t'+n}$ in K',

$$\begin{aligned}r\left(\mathbf{a}^{\alpha}_{t+n}\left(s^{t+n-1}\right), s_{t+n}\right) &= r\left(\alpha\left(\mu\left(\cdot|\mathbf{h}^{\alpha}_{t+n}(s^{t+n-1})\right)\right), k_n\right)\\ &= r\left(\alpha\left(\mu\left(\cdot|\mathbf{h}^{\alpha}_{t'+n}(s^{t'+n-1})\right)\right), k_n\right)\\ &= r\left(\mathbf{a}^{\alpha}_{t'+n}\left(s^{t'+n-1}\right), s_{t'+n}\right).\end{aligned} \tag{7}$$

Now, we restart to explicitly highlight the dependence on $(k_0, ..., k_n)$ of K. Moreover, for every $n \in \mathbb{N}_0$ and for every $(k_1, ..., k_n) \in S^n$, let $r(k_0, ..., k_n) = r\left(\mathbf{a}^{\alpha}_{t+n}\left(s^{t+n-1}\right), s_{t+n}\right) = r\left(\mathbf{a}^{\alpha}_{t'+n}\left(s^{t'+n-1}\right), s_{t'+n}\right)$, where $s^{t+n} \in K(k_0, ..., k_n)$ and $s^{t'+n} \in K'(k_0, ..., k_n)$. By (7), this quantity is well defined. We have:

$$\begin{aligned}&\sum_{s^{t+n}\in S^{t+n}} r\left(\mathbf{a}^{\alpha}_{t+n}\left(s^{t+n-1}\right), s_{t+n}\right) p_{\theta}(s^{t+n}|h_t)\\ &= \sum_{s^{t+n}\in S^{t+n}} r\left(\mathbf{a}^{\alpha}_{t+n}\left(s^{t+n-1}\right), s_{t+n}\right)\prod_{i=0}^{n}\theta(k_n)\\ &= \sum_{s^{t'+n}\in S^{t'+n}} r\left(\mathbf{a}^{\alpha}_{t'+n}\left(s^{t'+n-1}\right), s_{t'+n}\right) p_{\theta}(s^{t'+n}|h'_{t'}).\end{aligned}$$

Finally, since we have $\mu(\cdot \mid h_t) = \mu(\cdot \mid h'_{t'})$, this implies that:

$$\phi^{-1}\left(\int_{\operatorname{supp}\mu(\cdot|h_t)} \phi\left(\sum_{s^{t+n}\in S^{t+n}} r\left(\mathbf{a}^{\alpha}_{t+n}\left(s^{t+n-1}\right), s_{t+n}\right) p_\theta(s^{t+n}|h_t)\right)\mu\left(\mathrm{d}\theta \mid h_t\right)\right)$$

$$= \phi^{-1}\left(\int_{\operatorname{supp}\mu(\cdot|h'_{t'})} \phi\left(\sum_{s^{t'+n}\in S^{t'+n}} r\left(\mathbf{a}^{\alpha}_{t'+n}\left(s^{t'+n-1}\right), s_{t'+n}\right) p_\theta(s^{t'+n}|h'_{t'})\right)\mu\left(\mathrm{d}\theta \mid h'_{t'}\right)\right)$$

and the thesis follows. □

Proof of Proposition 2. The result follows by considering the subgame perfect equilibrium of an ancillary game. We describe the structure of this game:

- The set of players is $P = \mathbb{N} \cup \{0\}$. That is, players are the different periods and Nature;
- The set of actions available to each player $t \in \mathbb{N}$ is A, the set of actions available to Nature is S;
- The timing is as follows: at period $t \in \mathbb{N}$, player t chooses an action. After player t has chosen his action, Nature chooses $s_t \in S$ accordingly to a (stationary) uniform distribution over the states;
- Every player observes the sequence of actions played and states realized in the previous periods.

Notice that the set of subgames where a player $t \in \mathbb{N}$ is going to move is one to one with the set H_t. Therefore, a strategy profile σ of the ancillary game is mapped in an obvious way into a strategy of our single-agent decision problem.

- Finally, we specify the payoff of player t when the strategy profile of the single-agent decision problem induced by σ is α and the sequence of state realized is s^∞:

$$U_t\left(\alpha, s^\infty\right) = V\left(\alpha, \mu\left(\cdot|\mathbf{h}^{\alpha}_t(s^t)\right)|\mathbf{h}^{\alpha}_t(s^t)\right)$$

  where V is the value in the single-agent decision problem.

The ancillary game thus obtained is a perfect information game, and by Lemma 4 it satisfies all the assumptions of Theorem 4 in Hellwig and Leininger (1987). Therefore it admits a subgame-perfect equilibrium in pure strategies α. Moreover, by Theorem 4 in Hellwig and Leininger (1987) and the coupling argument of Lemma 4 every player with the same belief plays the same action, and by definition of the payoff function of each player, (α, μ) is rational. □

A.3. Convergence to SCE

By Lemma 4, we know that if h_t and $h'_{t'}$ are two histories such that $p_\mu(I(h_t))$ and $p_\mu(I(h'_{t'}))$ are strictly positive, then $\mu(\cdot \mid h_t) = \mu(\cdot \mid h'_{t'})$ implies $V(\alpha, \mu|h_t) = V\left(\alpha, \mu|h'_{t'}\right)$. Therefore, the value function at a particular history depends on the history only through the updated beliefs. In particular, whenever $p_\mu(I(h_t)) > 0$,

$$V(\alpha, \mu | h_t) = V(\alpha, \mu(\cdot|h_t)|h_1).$$

Therefore, to ease notation, in the following proofs we will use the ancillary function $\hat{V}$ mapping stationary strategies and beliefs into real numbers:

$$\hat{V}(\alpha, \nu) := V(\alpha, \nu | h_1).$$

By the previous argument, $\hat{V}(\alpha, \nu) = V(\alpha, \mu | h_t)$ whenever $\nu = \mu(\cdot|h_t)$.

Lemma 10. *Let $\left(\alpha, \mu, \bar{\theta}\right)$ be such that:*

1. $\mu\left(\{\theta \in \Theta : p_\theta^\alpha = p_{\bar{\theta}}^\alpha\}\right) = 1$;
2. *For every action a, period t, and information history h_t,*

$$p_\mu(I(h_t)) > 0 \Rightarrow V(\alpha, \mu \mid h_t) \geq V(\alpha/(h_t, a), \mu \mid h_t).$$

Then, $\left(\alpha(\mu), \mu, \bar{\theta}\right)$ is an SCE.

Proof. It is immediate to see that condition (1) implies condition (i) of SCE. Now, we show that $\left(\alpha, \mu, \bar{\theta}\right)$ features myopic best reply on path, that is, $(\alpha(\mu), \mu, \overline{\theta})$ satisfies property (ii) of SCE. By way of contradiction, suppose there is $a \in A$ such that:

$$\phi^{-1}\left(\int_\Theta \phi\left(\sum_{s\in S} r(a,s)\theta(s)\right)\mu(\mathrm{d}\theta)\right) > \phi^{-1}\left(\int_\Theta \phi\left(\sum_{s\in S} r(\alpha(\mu),s)\theta(s)\right)\mu(\mathrm{d}\theta)\right).$$

By property 2, it must be the case that $\hat{V}(\alpha/a, \mu) \leq \hat{V}(\alpha, \mu)$, where α/a is the strategy that prescribes a in the first period to come and coincide with α otherwise. However, we have:

$$\begin{aligned}
&\hat{V}(\alpha, \mu)\\
&= \sum_{s\in S} r(\alpha(\mu), s)\overline{\theta}(s) + \delta \hat{V}(\alpha, \mu)\\
&\leq \sum_{s\in S} r(\alpha(\mu), s)\overline{\theta}(s) + \delta \min_{m: F\left(a,\theta_\mu\right)(m)>0} \hat{V}(\alpha, \mu(\cdot|(a,m)))\\
&< \phi^{-1}\left(\int_\Theta \phi\left(\sum_{s\in S} r(a,s)\theta(s)\right)\mu(\mathrm{d}\theta)\right) +\\
&\delta \min_{m: F\left(a,\theta_\mu\right)(m)>0} \left(\sum_{\tau=2}^{\infty} \delta^{\tau-2}\phi^{-1}\right.\\
&\left.\times\left(\int_\Theta \phi\left(\sum_{s^\tau \in S^\tau} r\left(\mathbf{a}_\tau^\alpha\left(s^{\tau-1}\right), s_\tau\right) p_\theta(s^\tau|(a,m))\right)\mu\left(\mathrm{d}\theta \mid (a,m)\right)\right)\right)\\
&= \phi^{-1}\left(\int_\Theta \phi\left(\sum_{s\in S} r(a,s)\theta(s)\right)\mu(\mathrm{d}\theta)\right) +\\
&\delta\left(\sum_{\tau=2}^{\infty} \delta^{\tau-2}\phi^{-1}\left(\min_{m: F\left(a,\theta_\mu\right)(m)>0}\right.\right.
\end{aligned}$$

$$\times \left(\int_{\Theta} \phi \left(\sum_{s^\tau \in S^\tau} r\left(\mathbf{a}_\tau^\alpha \left(s^{\tau-1} \right), s_\tau \right) p_\theta(s^\tau | (a,m)) \right) \mu\left(\mathrm{d}\theta \mid (a,m) \right) \right) \right) \Bigg)$$

$$\leq \phi^{-1} \left(\int_{\Theta} \phi \left(\sum_{s \in S} r(a,s)\theta(s) \right) \mu(\mathrm{d}\theta) \right) +$$

$$\delta \left(\sum_{\tau=2}^{\infty} \delta^{\tau-2} \phi^{-1} \left(\int_{\Theta} \phi \left(\sum_{s^\tau \in S^\tau} r\left(\mathbf{a}_\tau^\alpha \left(s^{\tau-1} \right), s_\tau \right) p_\theta(s^\tau | (a,m)) \right) \mathbb{E}_{p_\mu} \left[\mu(\mathrm{d}\theta | (a,m)) \right] \right) \right)$$

$$= \phi^{-1} \left(\int_{\Theta} \phi \left(\sum_{s \in S} r(a,s)\theta(s) \right) \mu(\mathrm{d}\theta) \right) +$$

$$\delta \left(\sum_{\tau=2}^{\infty} \delta^{\tau-2} \phi^{-1} \left(\int_{\Theta} \phi \left(\sum_{s^\tau \in S^\tau} r\left(\mathbf{a}_\tau^\alpha \left(s^{\tau-1} \right), s_\tau \right) p_\theta(s^\tau | (a,m)) \right) \mu\left(\mathrm{d}\theta \right) \right) \right)$$

$$= \sum_{\tau=1}^{\infty} \delta^{\tau-2} \phi^{-1} \left(\int_{\Theta} \phi \left(\sum_{s^\tau \in S^\tau} r\left(\mathbf{a}_\tau^{\alpha/a} \left(s^{\tau-1} \right), s_\tau \right) p_\theta(s^\tau | (a,m)) \right) \mu\left(\mathrm{d}\theta \right) \right)$$

$$= V(\alpha/a, \mu),$$

where the first equality comes from property 1, the strict inequality comes from hypothesis, the second equality comes from the fact that ϕ is strictly increasing, the third equality by Lemma 6, and the fourth and fifth equalities by the definition of α/a. Note that we will be done as soon as we prove the first weak inequality, that is:

$$\hat{V}(\alpha, \mu) \leq \min_{m : F\left(a, \theta_\mu\right)(m) > 0} \hat{V}(\alpha, \mu(\cdot|(a,m)).$$

Indeed, it would follow that $\hat{V}(\alpha, \mu) < \hat{V}(\alpha/a, \mu)$, a contradiction with the fact that $(\alpha, \mu, \bar{\theta})$ satisfies 2.

Suppose that there exists m such that $F\left(a, \theta_\mu\right)(m) > 0$ with $\hat{V}(\alpha, \mu(\cdot|(a,m))) < \hat{V}(\alpha, \mu)$. The fact that $F\left(a, \theta_\mu\right)(m) > 0$ implies that $\theta_\mu(I(a,m)) > 0$. On the other hand, by property 1, $\mu\left(\left\{ \theta \in \Theta : p_\theta^\alpha = p_{\bar{\theta}}^\alpha \right\} \right) = 1$, and in particular:

$$\mu\left(\left\{ \theta \in \Theta : F\left(\alpha(\mu), \theta\right)(m) = F\left(\alpha(\mu), \bar{\theta}\right)(m) \right\} \right) = 1.$$

Then, let $B = \left\{ \theta \in \Theta : F\left(\alpha(\mu), \theta\right) = F\left(\alpha(\mu), \bar{\theta}\right) \right\}$. By Bayes rule, we have that:

$$\mu\left(B | (a,m)\right) = \mu\left(B\right) = 1.$$

But then, it follows that:

$$\hat{V}(\alpha, \mu(\cdot|(a,m)))$$
$$< (1-\delta)\hat{V}(\alpha, \mu) + \delta \hat{V}(\alpha, \mu(\cdot|(a,m))$$
$$= \phi^{-1} \left(\int_{\Theta} \phi \left(\sum_{s \in S} r(\alpha(\mu), s)\theta(s) \right) \mu(\mathrm{d}\theta) \right) + \delta \hat{V}(\alpha, \mu(\cdot|(a,m))$$

$$= \phi^{-1}\left(\int_{\Theta}\phi\left(\sum_{s\in S} r(\alpha(\mu),s)\theta(s)\right)\mu(\mathrm{d}\theta|(a,m))\right) + \delta\hat{V}(\alpha,\mu(\cdot|(a,m))$$

$$= \sum_{\tau=1}^{\infty}\delta^{\tau-1}\phi^{-1}\left(\int_{\Theta}\phi\left(\sum_{s^\tau\in S^\tau} r\left(\mathbf{a}_\tau^{\alpha/\alpha(\mu)}\left(s^{\tau-1}\right),s_\tau\right)p_\theta(s^\tau)\right)\mu\left(\mathrm{d}\theta|(a,m)\right)\right)$$

$$= \hat{V}(\alpha/(\alpha(\mu),\mu(\cdot|(a,m))).$$

This contradicts the fact that $(\alpha,\mu,\overline{\theta})$ satisfies property 2. □

Proof of Proposition 3. Let $\left(\alpha,\mu,\bar{\theta}\right)$ be consistent from period T. First, we have that the hypotheses of Lemma 7 are satisfied, so let E be as in the corresponding proof. Define

$$\hat{E} = E\bigcap_{t\in\mathbb{N}}\left\{s^\infty : \bar{\theta}\left(\iota_t^\alpha\left(s^{t-1}\right)\right) > 0\right\}.$$

By Lemma 3, the value (4) is continuous in beliefs μ. Fix $s^\infty\in\hat{E}$; for every a in A,

$$\lim_{t\to\infty}\hat{V}(\alpha/a,\mu(\cdot|\mathbf{h}_t^\alpha\left(s^{t-1}\right))) = \hat{V}(\alpha/a,\mu_{s^\infty}^\alpha).$$

Let $A_\infty := \arg\max_{a\in A}\hat{V}(\alpha/a,\mu_{s^\infty}^\alpha)$. Note that, in general, our definition of rationality does not require that $\alpha(\mu_{s^\infty}^\alpha)\in A_\infty$. Indeed, if there is no h_t such that $p_\mu(I(h_t))>0$ and $\mu(\cdot|h_t)=\mu_{s^\infty}^\alpha$, then $\alpha\left(\mu_{s^\infty}^\alpha\right)$ does not need to satisfy the one-deviation property. Since $s^\infty\in\hat{E}$, it follows that $p_{\overline{\theta}}(I(\mathbf{h}_t^\alpha\left(s^{t-1}\right)))>0$ for every finite t. By Assumption 2, $p_\mu(I(\mathbf{h}_t^\alpha\left(s^{t-1}\right)))>0$. Hence,

$$\alpha(\mu(\cdot|\mathbf{h}_t^\alpha\left(s^{t-1}\right)))\in\arg\max_{a\in A}\hat{V}(\alpha/a,\mu(\cdot|\mathbf{h}_t^\alpha\left(s^{t-1}\right))).$$

Now, let $a\notin A_\infty$, and fix $a^*\in A_\infty$. We have that

$$\begin{aligned}\lim_{t\to\infty}\hat{V}(\alpha/a,\mu(\cdot|\mathbf{h}_t^\alpha\left(s^{t-1}\right))) &= \hat{V}(\alpha/a,\mu_{s^\infty}^\alpha)\\ &< \max_{a'\in A}\hat{V}(\alpha/a',\mu_{s^\infty}^\alpha) = \hat{V}(\alpha/a^*,\mu_{s^\infty}^\alpha)\\ &= \lim_{t\to\infty}\hat{V}(\alpha/a^*,\mu(\cdot|\mathbf{h}_t^\alpha\left(s^{t-1}\right))).\end{aligned}$$

Hence there exists $T_{s^\infty}^a$ such that $a\notin\alpha(\mu(\cdot|\mathbf{h}_t^\alpha\left(s^{t-1}\right)))$ for every $t\geq T_{s^\infty}^a$. Let $T_{s^\infty}^* = \max_{a\in A\setminus A_\infty}T_{s^\infty}^a$. Then, from $T_{s^\infty}^*$ onward, the only actions played are in A_∞, that is, they satisfy the one-deviation property with respect to the limit beliefs $\mu_{s^\infty}^\alpha$. Let $\widehat{T}_{s^\infty}=\max\left\{T,T_{s^\infty}^*\right\}$; we have that from $\widehat{T}_{s^\infty}$ onward the action prescribed by strategy α, $\mathbf{a}_t^\alpha(s^{t-1})$, satisfies the one-deviation property with respect to beliefs $\mu_{s^\infty}^\alpha$, and $\mu_{s^\infty}^\alpha$ is confirmed given such action. By Lemma 10, this implies that $(\mathbf{a}_t^\alpha(s^{t-1}),\mu_{s^\infty}^\alpha,\overline{\theta})$ is an SCE for every $t\geq\widehat{T}_{s^\infty}$. □

Proof of Proposition 4. By hypothesis, we know that $\left(\mathbf{a}_t^\alpha(s^\infty),\mu\left(\cdot|\mathbf{h}_t^\alpha\left(s^\infty\right)\right)\right)$ converges to an SCE. Therefore, there exists $\widehat{T}$ such that, for every $t\geq\widehat{T}$, the triple $(\mathbf{a}_t^\alpha(s^{t-1}),\mu_{s^\infty}^\alpha,\overline{\theta})$ is an SCE, and so:

$$\mathbf{a}_t^\alpha(s^{t-1})\in\arg\max_{a\in A}\phi^{-1}\left(\int_{\Theta}\phi\left(R\left(a,\theta\right)\right)\mu_{s^\infty}^\alpha(\mathrm{d}\theta)\right) = \left\{a^*\right\}.$$

It follows that $(a^*, \mu^{\alpha}_{s^\infty}, \overline{\theta})$ is an SCE. Now, let $a \neq a^*$. By Lemmata 5 and 7,

$$\begin{aligned}
&\lim_{t\to\infty} \phi^{-1}\left(\int_\Theta \phi\left(\sum_{s\in S} r(a^*,s)\theta(s)\right)\mu(\mathrm{d}\theta|\mathbf{h}^{\alpha}_t\left(s^{t-1}\right))\right)\\
&= \phi^{-1}\left(\int_\Theta \phi\left(\sum_{s\in S} r(a^*,s)\theta(s)\right)\mu^{\alpha}_{s^\infty}(\mathrm{d}\theta)\right)\\
&> \phi^{-1}\left(\int_\Theta \phi\left(\sum_{s\in S} r(a,s)\theta(s)\right)\mu^{\alpha}_{s^\infty}(\mathrm{d}\theta)\right)\\
&= \lim_{t\to\infty} \phi^{-1}\left(\int_\Theta \phi\left(\sum_{s\in S} r(a,s)\theta(s)\right)\mu(\mathrm{d}\theta|\mathbf{h}^{\alpha}_t\left(s^{t-1}\right))\right).
\end{aligned}$$

Thus, there exists $\bar{T}_{a,s^\infty} > \widehat{T}$ such that $t > \bar{T}_{a,s^\infty}$ implies:

$$\begin{aligned}
&\phi^{-1}\left(\int_\Theta \phi\left(\sum_{s\in S} r(a^*,s)\theta(s)\right)\mu(\mathrm{d}\theta|\mathbf{h}^{\alpha}_t\left(s^{t-1}\right))\right)\\
&> \phi^{-1}\left(\int_\Theta \phi\left(\sum_{s\in S} r(a,s)\theta(s)\right)\mu(\mathrm{d}\theta|\mathbf{h}^{\alpha}_t\left(s^{t-1}\right))\right).
\end{aligned}$$

Let $\bar{T}_{s^\infty} = \max_{a\in A\setminus a^*} \bar{T}_{a,s^\infty}$. We have that $t > \bar{T}_{s^\infty}$ implies

$$\begin{aligned}
&\phi^{-1}\left(\int_\Theta \phi\left(\sum_{s\in S} r(a^*,s)\theta(s)\right)\mu(\mathrm{d}\theta|\mathbf{h}^{\alpha}_t\left(s^{t-1}\right))\right)\\
&= \max_{a\in A} \phi^{-1}\left(\int_\Theta \phi\left(\sum_{s\in S} r(a,s)\theta(s)\right)\mu(\mathrm{d}\theta|\mathbf{h}^{\alpha}_t\left(s^{t-1}\right))\right).
\end{aligned}$$

The thesis follows. □

Proof of Corollary 3. By Proposition 3, there exists a set $E \subseteq S^\infty$ with $p_{\bar{\theta}}(E) = 1$ such that convergence to an SCE happens on that set. By Proposition 1, there exists a set $E^* \subseteq S^\infty$ with $p_{\bar{\theta}}(E^*) = 1$ such that, if $s^\infty \in E^*$, then for every $a \in \mathbf{a}^{\alpha}_\infty(s^\infty)$:

$$\lim_{t\to\infty} \mu(\left\{\theta\in\Theta : F(a,\theta) = F\left(a,\bar{\theta}\right)\right\} \mid \mathbf{h}^{\alpha}_t\left(s^\infty\right)) = 1. \tag{8}$$

Let $G = E \cap E^*$. Own-action independence and 8 imply that, if $s^\infty \in G$, then for every $\hat{a} \in A$:

$$\lim_{t\to\infty} \mu(\left\{\theta\in\Theta : F\left(\hat{a},\theta\right) = F\left(\hat{a},\bar{\theta}\right)\right\} \mid \mathbf{h}^{\alpha}_t\left(s^\infty\right)) = 1.$$

Thus, for every $s^\infty \in G \subseteq E^*$,

$$\phi^{-1}\left(\int_{\Theta}\phi\left(R\left(\hat{a},\theta\right)\right)\mu_{s^{\infty}}^{\alpha}(\mathrm{d}\theta)\right)=R\left(\hat{a},\bar{\theta}\right); \tag{9}$$

that is, the value of each action under the limit belief is equal to the objective value. Since $s^{\infty}\in G\subseteq E$, there exists a finite time t such that $(\mathbf{a}_{\tau}^{\alpha}(s^{\tau-1}),\mu_{s^{\infty}}^{\alpha},\bar{\theta})$, $\tau\geq t$, forms an SCE. By definition of SCE and Equation (9), this means that for $\tau\geq t$ only the objective myopic best reply is played. □

A.4. Comparative dynamics

Since this section deals with different levels of ambiguity attitude, we make the dependence of the value on the ambiguity attitude explicit by writing $\hat{V}_{\phi}(\alpha,\mu)$ in place of $\hat{V}(\alpha,\mu)$.

Proof of Proposition 5. Since (α,μ) is rational under ambiguity neutrality, it satisfies the one-deviation property for every h_t such that $p_{\mu}(I(h_t))>0$ and $\nu=\mu(\cdot|h_t)$:

$$V_{\mathrm{Id}}(\alpha,\mu|h_t)\geq V_{\mathrm{Id}}(\alpha/(h_t,a),\mu|h_t)$$

for every $a\in A$. Since the spaces of action and state are finite and $\delta<1$, our problem is continuous at infinity. By the one-deviation principle for ambiguity neutral agents (see e.g., Theorem 4.2 in Fudenberg and Tirole, 1991), this implies that for every alternative strategy γ and every history h_{τ} with $p_{\mu}(I(h_t))>0$:

$$V_{\mathrm{Id}}(\alpha,\mu|h_t)\geq V_{\mathrm{Id}}(\gamma,\mu|h_t). \tag{10}$$

Let ν be in the belief range of μ, say $\nu=\mu(\cdot|h_t)$ and suppose, by way of contradiction, that $\beta(\nu)$ prescribes an ambiguous action while $\alpha(\nu)$ is unambiguous given ν. Then,

$$\begin{aligned}V_{\phi}(\beta,\mu|h_t)&\geq V_{\phi}(\beta/(\alpha(\nu),h_t),\mu|h_t)\\&=R(\alpha(\nu),p_{\nu})+\delta\sum_{m\in M}F(\alpha(\nu),p_{\nu})(m)\,V_{\phi}(\beta,\mu|(h_t,(\alpha(\nu),m)))\\&=R(\alpha(\nu),p_{\nu})+\delta V_{\phi}(\beta,\mu|h_t)\end{aligned}$$

where the inequality follows from the one-deviation property, the second equality follows from the assumption that $\alpha(\nu)$ is unambiguous, and the third equality follows from $\alpha(\nu)$ being ν-unambiguous and Lemma 4. Therefore,

$$V_{\phi}(\beta,\mu|h_t)\geq\frac{R(\alpha(\nu),p_{\nu})}{1-\delta}.$$

Since $\alpha(\nu)$ is ν-unambiguous,

$$\begin{aligned}V_{\phi}(\alpha,\mu|h_t)&=R(\alpha(\nu),p_{\nu})+\delta\sum_{m\in M}F(\alpha(\nu),p_{\nu})(m)\,V_{\phi}(\alpha,\mu|(h_t,(\alpha(\nu),m)))\\&=\sum_{k=0}^{\infty}\delta^{k}R(\alpha(\nu),p_{\nu})=\frac{R(\alpha(\nu),p_{\nu})}{1-\delta}.\end{aligned}$$

Then by (10), and Jensen's inequality (see page 294 in Billingsley, 2012 for the version used here),

$$\frac{R\left(\alpha\left(\nu\right),p_{\nu}\right)}{1-\delta}=V_{\mathrm{Id}}\left(\alpha,\nu|h_{t}\right)\geq V_{\mathrm{Id}}\left(\beta,\nu|h_{t}\right)>V_{\phi}\left(\beta,\nu|h_{t}\right)\geq\frac{R\left(\alpha\left(\nu\right),p_{\nu}\right)}{1-\delta},$$

a contradiction. □

Proof of Proposition 6. Let $p_{\bar{\theta}}\left(\mathbf{h}_{t}^{\alpha}\left(s^{t-1}\right)\right)>0$, and let $\left(\alpha\left(\mu\left(\cdot|\mathbf{h}_{t}^{\alpha}\left(s^{t-1}\right)\right)\right),\mu\left(\cdot|\mathbf{h}_{t}^{\alpha}\left(s^{t-1}\right)\right),\bar{\theta}\right)$ be an SCE under ambiguity neutrality. By Assumption 2, an action that is unambiguous given μ will be unambiguous given $\mu\left(\cdot|\mathbf{h}_{t}^{\alpha}\left(s^{t-1}\right)\right)$. Therefore, when an action $a\neq a^{*}$ is chosen, with probability 1 it does not induce any updating in beliefs. Then, stationarity of the strategy implies that the same action a will be chosen in the following period. Thus, the sequence $\left(\alpha\left(\mu\left(\cdot|\mathbf{h}_{\tau}^{\alpha}\left(s^{\tau-1}\right)\right)\right)\right)_{\tau=1}^{t}$ has almost surely the form $(a^{*},...,a^{*},a,...,a)$ for some $a\in A$.[35]

The same reasoning guarantees that $\left(\beta\left(\mu\left(\cdot|\mathbf{h}_{\tau}^{\beta}\left(s^{\tau-1}\right)\right)\right)\right)_{\tau=1}^{t}$ also has almost surely the form $(a^{*},...,a^{*},a,...,a)$ for some $a\in A$. We show that for almost every s^{∞}, the sequence of a^{*} is (weakly) shorter under β. In particular, suppose that $t\in\mathbb{N}$, $\bar{\theta}\left(s^{t-1}\right)>0$, and let:

$$\left(\alpha\left(\mu\left(\cdot|\mathbf{h}_{\tau}^{\alpha}\left(s^{\tau-1}\right)\right)\right)\right)_{\tau=1}^{t}=\left(a^{*},...,a^{*}\right)=\left(\beta\left(\mu\left(\cdot|\mathbf{h}_{\tau}^{\beta}\left(s^{\tau-1}\right)\right)\right)\right)_{\tau=1}^{t}.$$

Then,

$$\mathbf{h}_{t}^{\alpha}\left(s^{t-1}\right)=\left(a^{*},f\left(a^{*},s_{\tau}\right)\right)_{\tau=1}^{t-1}=\mathbf{h}_{t}^{\beta}\left(s^{t-1}\right),$$

and by Assumption 2:

$$\mu\left(\cdot|\mathbf{h}_{\tau}^{\alpha}\left(s^{\tau-1}\right)\right)=\mu\left(\cdot|\mathbf{h}_{t}^{\beta}\left(s^{t-1}\right)\right).$$

By Proposition 5, $\alpha\left(\mu\left(\cdot|\mathbf{h}_{t}^{\alpha}\left(s^{t-1}\right)\right)\right)\neq a^{*}$ implies $\beta\left(\mu\left(\cdot|\mathbf{h}_{t}^{\alpha}\left(s^{t-1}\right)\right)\right)\neq a^{*}$; with probability 1, if strategy α ends experimentation after $t-1$ periods, β ends experimentation in at most $t-1$ periods. Therefore, $\left(\beta\left(\mu\left(\cdot|\mathbf{h}_{\tau}^{\beta}\left(s^{\tau-1}\right)\right)\right)\right)_{\tau=1}^{t}$ has the form $(a^{*},...,a^{*},a,...,a)$ with a (weakly) shorter sequence of a^{*}. If $\beta\left(\mu\left(\cdot|\mathbf{h}_{t}^{\beta}\left(s^{t-1}\right)\right)\right)\neq a^{*}$, $\left(\beta\left(\mu\left(\cdot|\mathbf{h}_{t}^{\beta}\left(s^{t}\right)\right)\right),\mu\left(\cdot|\mathbf{h}_{t}^{\beta}\left(s^{t}\right)\right),\bar{\theta}\right)$ is an SCE by Lemma 10 and rationality of (β,μ). If $\left(\beta\left(\mu\left(\cdot|\mathbf{h}_{\tau}^{\beta}\left(s^{\tau-1}\right)\right)\right)\right)_{\tau=1}^{t}=\left(a^{*}\right)_{\tau=1}^{t}$, then α and β have prescribed the same action a^{*} at every node, and therefore:

$$\begin{aligned}&\left(\alpha\left(\mu\left(\cdot|\mathbf{h}_{t}^{\alpha}\left(s^{t-1}\right)\right)\right),\mu\left(\cdot|\mathbf{h}_{t}^{\alpha}\left(s^{t-1}\right)\right),\bar{\theta}\right)\\&\quad=\left(\beta\left(\mu\left(\cdot|\mathbf{h}_{t}^{\beta}\left(s^{t-1}\right)\right)\right),\mu\left(\cdot|\mathbf{h}_{t}^{\beta}\left(s^{t-1}\right)\right),\bar{\theta}\right).\end{aligned}$$

Again, by Lemma 10, the fact that the L.H.S. is an SCE with ambiguity neutrality, and the definition of SCE, the R.H.S. is an SCE under ambiguity aversion. □

Proof of Proposition 7. First, suppose that the objectively-optimal action played in the SCE is a^{*}. Then, given the original belief μ and infinite history s^{∞}, $\left(\mathbf{a}_{t}^{\beta},\mu\left(\cdot|\mathbf{h}_{t}^{\beta}\right)\right)$ converges to an SCE $\left(a^{*},\mu^{*},\bar{\theta}\right)$ if and only if $\beta\left(\mu\left(\cdot|\mathbf{h}_{\tau}^{\beta}\left(s^{\infty}\right)\right)\right)=a^{*}$ for every τ. But Proposition 5 guarantees that this can happen only if $\alpha\left(\mu\left(\cdot|\mathbf{h}_{\tau}^{\alpha}\left(s^{\infty}\right)\right)\right)=a^{*}$ for every τ, and therefore $\left(\mathbf{a}_{\tau}^{\alpha},\mu\left(\cdot|\mathbf{h}_{\tau}^{\alpha}\right)\right)$ also converges to $\left(a^{*},\mu^{*},\bar{\theta}\right)$.

[35] Possibly including the cases $\left(a^{*},...,a^{*}\right)$ and $(a,...,a)$.

Second, suppose that the objectively-optimal action is $a \neq a^*$. Consider s^∞ where convergence of $\left(\mathbf{a}_t^\beta, \mu\left(\cdot|\mathbf{h}_t^\beta\right)\right)$ to an SCE $\left(a, \nu, \bar{\theta}\right)$ happens. Then, by definition of SCE, $R\left(a,\theta\right) = R\left(a,\bar{\theta}\right)$ is constant on $\operatorname{supp}\nu$. At the same time, Proposition 3 guarantees convergence of $\left(\mathbf{a}_t^\alpha, \mu\left(\cdot|\mathbf{h}_t^\alpha\right)\right)$ to an SCE $\left(\hat{a}, \hat{\mu}, \bar{\theta}\right)$. Assumption 2 implies that $\operatorname{supp}\mu\left(\cdot|\mathbf{h}_t^a\right) \subset \operatorname{supp}\mu\left(\cdot\right)$ for every h_t and therefore $R\left(a,\theta\right) = R\left(a,\bar{\theta}\right)$ is constant on $\operatorname{supp}\hat{\mu}$. Since, by definition of SCE, $R\left(\hat{a},\theta\right) = R\left(\hat{a},\bar{\theta}\right)$ is also constant on $\operatorname{supp}\hat{\mu}$ and:

$$\hat{a} \in \arg\max_{a' \in A} \phi^{-1}\left(\int_\Theta \phi\left(\sum_{s\in S} r(a',s)\theta(s)\right)\mu^*(\mathrm{d}\theta)\right),$$

we must have $R\left(\hat{a},\bar{\theta}\right) \geq R\left(a,\bar{\theta}\right) = \max_{a'\in A} R\left(a',\bar{\theta}\right)$. □

Proof of Proposition 8. Let $\left(\bar{a},\mu,\bar{\theta}\right)$ be an SCE under ambiguity attitude ϕ. The optimality condition for an SCE gives:

$$\bar{a} \in \arg\max_{a' \in A} \phi^{-1}\left(\int_\Theta \phi\left(\sum_{s\in S} r\left(a',s\right)\theta\left(s\right)\right)\mu\left(\mathrm{d}\theta\right)\right). \tag{11}$$

Therefore, there exists θ^* in $\operatorname{supp}\mu$ with:

$$\sum_{s\in S} r\left(\bar{a},s\right)\theta^*\left(s\right) \geq \sum_{s\in S} r\left(a^*,s\right)\theta^*\left(s\right).$$

Next, consider an arbitrary $a \neq a^*$; such actions are unambiguous given μ. By (11):

$$\begin{aligned}\sum_{s\in S} r\left(\bar{a},s\right)\theta^*\left(s\right) &= \phi^{-1}\left(\int_\Theta \phi\left(\sum_{s\in S} r\left(\bar{a},s\right)\theta\left(s\right)\right)\mu\left(\mathrm{d}\theta\right)\right)\\ &\geq \phi^{-1}\left(\int_\Theta \phi\left(\sum_{s\in S} r\left(a,s\right)\theta\left(s\right)\right)\mu\left(\mathrm{d}\theta\right)\right)\\ &= \sum_{s\in S} r\left(a,s\right)\theta^*\left(s\right).\end{aligned}$$

We have established that, for every $\sigma \in \Delta\left(A\right)$, $\sum_{s\in S} r\left(\bar{a},s\right)\theta^*\left(s\right) \geq \sum_{a'\in A}\sigma\left(a'\right) \times \sum_{s\in S} r(a',s)\theta^*(s)$. Now, define the function $\bar{v} : \Delta\left(A\right) \times \Delta(\operatorname{supp}\mu) \to \mathbb{R}$ as $\bar{v}(\sigma,\nu) = \bar{V}_{\mathrm{Id}}(\bar{a},\nu) - \sum_{a'\in A}\sigma\left(a'\right)\bar{V}_{\mathrm{Id}}(a',\nu)$. Then, for every $\sigma \in \Delta\left(A\right)$, $\bar{v}(\sigma,\delta_{\theta^*}) \geq 0$, and so:

$$\min_{\sigma\in\Delta(A)} \max_{\nu\in\Delta(\operatorname{supp}\mu)} \bar{v}(\sigma,\nu) \geq 0.$$

By Sion's minimax theorem in Sion (1958),

$$\max_{\nu\in\Delta(\operatorname{supp}\mu)} \min_{\sigma\in\Delta(A)} \bar{v}(\sigma,\nu) = \min_{\sigma\in\Delta(A)} \max_{\nu\in\Delta(\operatorname{supp}\mu)} \bar{v}(\sigma,\nu) \geq 0.$$

The function $\bar{v}$ is continuous and its domain is compact, so the set of maximizers is non-empty.[36] Pick any:

[36] Compactness of the domain of $\bar{v}$ follows from the fact that $\operatorname{supp}\mu\left(\cdot|h_t\right)$ is a closed subset of Δ.

$$\mu' \in \arg \max_{\nu \in \Delta(\operatorname{supp} \mu)} \left(\min_{\sigma \in \Delta(A)} \bar{v}(\sigma, \nu) \right).$$

We have, for every $a' \in A$,

$$\bar{V}_{\mathrm{Id}}(\bar{a}, \mu') - \bar{V}_{\mathrm{Id}}(a', \mu') \geq \min_{a' \in A} \bar{v}(\delta_{a'}, \mu') \geq 0.$$

It follows that $(\bar{a}, \mu', \bar{\theta})$ is an SCE under ambiguity neutrality. By Theorem 1 in Battigalli et al. (2015), $\left(\bar{a}, \mu', \bar{\theta}\right)$ is an SCE under ambiguity attitude ϕ'. □

Proof of Proposition 9. Since (α, μ) is rational under ϕ, and the DM is myopic, the one-deviation property reads:

$$\forall a \in A, \phi^{-1}\left(\int_{\Theta} \phi\left(R\left(\alpha(\nu), \hat{\theta}\right)\right) \nu\left(\mathrm{d}\hat{\theta}\right)\right) \geq \phi^{-1}\left(\int_{\Theta} \phi\left(R\left(a, \hat{\theta}\right)\right) \nu\left(\mathrm{d}\hat{\theta}\right)\right),$$

where ν is as in the proof of Proposition 5. Let (β, μ) be rational under ϕ'. Suppose, by way of contradiction, that $\beta(\nu)$ prescribes an ambiguous action. Then,

$$\begin{aligned}
\hat{V}_{\phi'}(\beta, \nu) &= \left(\phi'\right)^{-1}\left(\int_{\Theta} \phi'\left(R\left(\beta(\nu), \hat{\theta}\right)\right) \nu\left(\mathrm{d}\hat{\theta}\right)\right) \\
&\geq \left(\phi'\right)^{-1}\left(\int_{\Theta} \phi'\left(R\left(\alpha(\nu), \hat{\theta}\right)\right) \nu\left(\mathrm{d}\hat{\theta}\right)\right) \\
&= R\left(\alpha(\nu), p_{\nu}\right),
\end{aligned}$$

where the inequality follows from the one-deviation property, while the second equality follows from the assumption that $\alpha(\nu)$ is unambiguous. Therefore,

$$\hat{V}_{\phi'}(\beta, \nu) \geq R\left(\alpha(\nu), p_{\nu}\right).$$

By Jensen's inequality, since ϕ' is a strictly concave transformation of ϕ,

$$R\left(\alpha(\nu), p_{\nu}\right) = \hat{V}_{\phi}(\alpha, \nu) \geq \hat{V}_{\phi}(\beta, \nu) > \hat{V}_{\phi'}(\beta, \nu) \geq R\left(\alpha(\nu), p_{\nu}\right),$$

a contradiction. □

Proof of Proposition 10. The proof is very similar to that of Proposition 6, invoking Proposition 9 instead of Proposition 5. Further details are omitted. □

Proof of Proposition 11. As with the previous proposition, invoke Proposition 9 instead of Proposition 5. □

References

Anderson, C., 2012. Ambiguity aversion in multi-armed bandit problems. Theory Decis. 72, 15–33.
Arrow, K.J., 1971. Essays in the Theory of Risk-Bearing. Markham.
Arrow, K.J., Green, J.R., 1973. Notes on Expectations Equilibria in Bayesian Settings. Institute for Mathematical Studies in the Social Sciences. Working Paper 33.

Battigalli, P., Catonini, E., Lanzani, G., Marinacci, M., 2019. Ambiguity attitudes and self-confirming equilibrium in sequential games. Games Econ. Behav. 115, 1–29.
Battigalli, P., Cerreia-Vioglio, S., Maccheroni, F., Marinacci, M., 2015. Self-confirming equilibrium and model uncertainty. Am. Econ. Rev. 105, 646–677.
Billingsley, P., 2012. Probability and Measure. John Wiley & Sons, Ltd.
Blanchard, O., 1985. Debt, deficit, and finite horizon. J. Polit. Econ. 93, 223–247.
Cerreia-Vioglio, S., Maccheroni, F., Marinacci, M., Montrucchio, L., 2013a. Ambiguity and robust statistics. J. Econ. Theory 148, 974–1049.
Cerreia-Vioglio, S., Maccheroni, F., Marinacci, M., Montrucchio, L., 2013b. Classical subjective expected utility. Proc. Natl. Acad. Sci. USA 110, 6754–6759.
Doob, J.L., 1949. Application of the theory of martingales. Colloq. Int. Cent. Natl. Rech. Sci., 23–27.
Dynkin, E., 1965. Controlled random sequences. Theory Probab. Appl. 10, 1–14.
Easley, D., Kiefer, N.M., 1988. Controlling a stochastic process with unknown parameters. Econometrica 5, 1045–1064.
Epstein, L.G., Schneider, M., 2007. Learning under ambiguity. Rev. Econ. Stud. 74, 1275–1303.
Esponda, I., Pouzo, D., 2016. Berk–Nash equilibrium: a framework for modeling agents with misspecified models. Econometrica 84, 1093–1130.
Folland, G., 2013. Real Analysis: Modern Techniques and Their Applications. John Wiley & Sons, Ltd.
Fudenberg, D., He, S., 2018. Learning and type compatibility in signalling games. Econometrica 86, 1215–1255.
Fudenberg, D., Kreps, D.M., 1995. Learning in extensive-form games, I: self-confirming equilibria. Games Econ. Behav. 8, 20–55.
Fudenberg, D., Levine, D.K., 1993. Steady state learning and Nash equilibrium. Econometrica 61, 547–573.
Fudenberg, D., Levine, D.K., 1998. The Theory of Learning in Games. MIT Press, Cambridge, MA.
Fudenberg, D., Tirole, J., 1991. Game Theory. MIT Press, Cambridge, MA.
Gilboa, I., Schmeidler, D., 1989. Maxmin expected utility with a non-unique prior. J. Math. Econ. 18, 141–153.
Gittins, J.C., 1989. Multi-Armed Bandit Allocation Indices. Wiley-Interscience Series in Systems and Optimization. John Wiley & Sons, Ltd., Chichester.
Hanany, E., Klibanoff, P., 2009. Updating ambiguity averse preferences. B.E. J. Theor. Econ. 9.
Hanany, E., Klibanoff, P., Mukerji, S., 2019. Incomplete information games with ambiguity averse players. Am. Econ. J. Microecon. Forthcoming.
Hellwig, M., Leininger, W., 1987. On the existence of subgame-perfect equilibrium in infinite-action games of perfect information. J. Econ. Theory 43, 55–75.
Hinderer, K., 1970. Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter. Springer Science & Business Media.
Jewitt, I.E., Mukerji, S., 2017. Ordering ambiguous acts. J. Econ. Theory 171, 213–267.
Kalai, E., Lehrer, E., 1993. Rational learning leads to Nash equilibrium. Econometrica 1, 1019–1045.
Kalai, E., Lehrer, E., 1995. Subjective games and equilibria. Games Econ. Behav. 8, 123–163.
Klibanoff, P., Marinacci, M., Mukerji, S., 2005. A smooth model of decision making under ambiguity. Econometrica 73, 1849–1892.
Li, J., 2019. The k-armed bandit problem with multiple-priors. J. Math. Econ. 80, 22–38.
Maccheroni, F., Marinacci, M., Rustichini, A., 2006. Dynamic variational preferences. J. Econ. Theory 128, 4–44.
Marinacci, M., 2002. Learning from ambiguous urns. Stat. Pap. 43, 143–151.
Marinacci, M., 2015. Model uncertainty. J. Eur. Econ. Assoc. 13, 998–1076.
Rothschild, M., 1974. A two-armed bandit theory of market pricing. J. Econ. Theory 9, 185–202.
Sion, M., 1958. On general minimax theorems. Pac. J. Math. 171 (176).
Williams, D., 1991. Probability with Martingales. Cambridge University Press, Cambridge.