B. Supplementary Online Appendix

This document is a supplementary Online Appendix for the paper *Waizmann (2024) "AI in Action: Algorithmic Learning with Strategic Consumers"*, available here. References to equations, assumptions, theorems, and related elements pertain to that paper.

In this section, we elaborate on the main results of Section 3. We consider cases in which the short-run players' information about the algorithm's input is such that the algorithm is neither opaque nor transparent. We discuss under which condition the algorithm learns to play the Stackelberg action in each state.

Throughout, we maintain Assumptions 1 and 2. However, we do not impose Assumption 3. Given these assumptions, it is without loss of generality to assume that $|S_{\mathrm{SR}}| = 1$; otherwise, all statements hold conditional on each $s \in S_{\mathrm{SR}}$. Moreover, we assume that the signalling structure (Φ, p) has full support according to Definition 5.

Assume the short-run players can condition their actions on a subset of the outcomes of past interactions. Formally, for any period t let P^t be a partition of the set of histories

$$(S_Q \times A_Q \times U_Q)^t \, .$$

A strategy for the short-run player in period t, SR^t is a map

$$\sigma_{\mathrm{SR}^t} : P^{t-1} \times \Phi \to \Delta(A_{\mathrm{SR}}).$$

This formulation includes the special cases in which the short-run players observe the entire history up to period t as well as the special case in which the short-run players do not observe any outcome of previous interactions. Fix such a sequence $\{P^t\}$ of partitions throughout this section.

B.1. A necessary condition for learning the Stackelberg action

Consider the following auxiliary extensive-form game between the algorithmic player and one short-run player. The algorithmic player has $|S_Q|$-many types. Each cell $s \in S_Q$ corresponds to a type of the algorithmic player. Each type of the algorithmic player has the same set of actions A_Q. First, nature draws the algorithmic player's type s which is private information of the algorithmic player. After being informed of its type, the algorithmic player selects an action $a_Q \in A_Q$. After the algorithmic player takes its action a_Q, a signal $\phi \in \Phi$ is drawn according to $p(\cdot|a_Q)$. The short-run player observes the signal ϕ and takes an action $a_{\mathrm{SR}} \in A_{\mathrm{SR}}$. For each $a \in A_Q \times A_{\mathrm{SR}}$ the payoffs of the short-run player is given by $\mathbb{E}[u_{\mathrm{SR}}(a, \omega)]$ and the payoff of the algorithmic player is given by $\mathbb{E}[u_Q(a, \omega)|s]$ if its type is s. Denote this auxiliary game by $G(S_Q, (\Phi, p))$.

Let $\tau_Q : S_Q \to A_Q$ and $\tau_{\text{SR}} : \Phi \to A_{SR}$ be a strict Nash equilibrium of $G(S_Q, (\Phi, p))$.[60]

Theorem 5. *For any $\xi > 0$ there exists an open set of parameters $\langle Q^0, (\alpha^t), (\varepsilon^t)\rangle$ and for each (s, a_Q) an open neighborhood $\mathcal{O}_{(s,a_Q)}$ of*

$$\sum_{\phi\in\Phi} p(\phi \mid a_Q)\mathbb{E}[u_Q(a_Q, \tau_{\text{SR}}(\phi), \omega)|s]$$

such that if $Q^0(s, a_Q) \in \mathcal{O}_{(s,a_Q)}$ for each (s, a_Q), then in any equilibrium

$$\mathbb{P}\left[\lim_{T\to\infty} \frac{\sum_{t=0}^{T} \mathbb{1}\{(a_Q^t, s^t) = (\tau_Q(s), s)\}}{\sum_{t=0}^{T} \mathbb{1}\{s^t = s\}} = 1\right] \geq 1 - \xi.$$

Proof in Appendix B.3.1.

Corollary 2. *Suppose each action of the algorithmic player is the Stackelberg action for some state, i.e., for each a_Q there exists $s \in S_Q$ such that $a_Q = a_Q^{\text{Stack}}(s)$. For any $\xi > 0$ there exists $\overline{\gamma} > 0$ and an open set of parameters $\langle Q^0, (\alpha^t), (\varepsilon^t)\rangle$ such that for all γ-perfect signalling structures, $0 < \gamma < \overline{\gamma}$, and all states $s \in S_Q$,*

$$\mathbb{P}\left[\lim_{T\to\infty} \frac{\sum_{t=0}^{T} \mathbb{1}\{(a_Q^t, s^t) = (a_Q^{\text{Stack}}(s), s)\}}{\sum_{t=0}^{T} \mathbb{1}\{s^t = s\}} = 1\right] \geq 1 - \xi.$$

Theorem 5 is a generalization of Theorem 2. For the case of a transparent algorithm, Theorem 2 shows that, when the Q-values are close to the expected payoffs when the short-run player plays according to a Nash equilibrium of the auxiliary games, then with probability close to 1, both the algorithmic player and the short-run players play according to this Nash equilibrium in almost all future periods. Theorem 5 generalizes this result in two directions. First, it allows for the possibility that the short-run players do not observe the current state of the algorithm.[61] In this case, the auxiliary game is an extensive form game with the algorithmic player's type equal to the current state. Second, Theorem 5 allows for the possibility that the short-run players do not observe the outcome of past interactions. In particular, Theorem 5 admits the possibility that the short-run players have no information about the outcome of past interactions. The theorem thus applies to the case of an opaque algorithm as well.[62]

[60] The assumption that the signalling structure (Φ, p) has full support ensures that each Nash equilibrium of the auxiliary extensive-form game is outcome-equivalent to a sequential equilibrium.

[61] The case in which the short-run players observe the algorithm's state corresponds to the case $|S_Q| = 1$.

[62] This is the reason why the theorem includes a hypothesis on the algorithm's parameters.

			SR player	
		L	M	R
	T	$2, 2$	$0, 0$	$3, 0.4$
algorithm	M	$0.02, 0.5$	$1, 1$	$0.01, -0.02$
	B	$0.03, 0.9$	$4, 0.001$	$-0.01, 5$

Figure 5: The payoff matrix for Example 5. In this game, the unique Nash equilibrium and the Stackelberg outcome equal (T, L).

Why do Nash equilibria of the auxiliary game have a similar stability property when the algorithm is opaque? When the short-run players play according to a Nash equilibrium τ_{SR} of the auxiliary game in all periods, the environment is stationary. The algorithm then learns to play a best-response to τ_{SR}. This corresponds to playing according to the Nash equilibrium τ_Q. When experimentation rates are low and the algorithm's greedy actions correspond to playing according to the Nash equilibrium, the short-run players' best-response is τ_{SR}. When the short-run players know the Q-values, they keep playing according to τ_{SR} until the greedy action no longer corresponds to τ_Q. Theorems 2 and 5 show that the probability that this occurs for some period t can be made arbitrarily close to 1. When the short-run players do not know the Q-values – e.g., when they do not observe the outcome of past interactions – they play according to τ_{SR} as long as they believe with probability close to 1 that the greedy action corresponds to τ_Q.

Theorem 5 has an important consequence. Any condition on payoff functions that guarantees that the algorithm learns the Stackelberg action in each state must imply the following: the algorithmic player plays the Stackelberg action in any strict Nash equilibrium $(\tau_Q, \tau_{\text{SR}})$ of the auxiliary extensive-form game, i.e., $\tau_Q(s) = a_Q^{\text{Stack}}(s)$.[63] One might therefore conjecture that a sufficient condition for convergence to the Stackelberg outcome is that the Stackelberg outcome corresponds to a unique strict Nash equilibrium. The next example shows that this conjecture is incorrect.

Example 5. Suppose that $|S_Q| = 1$ and the signals ϕ are pure noise, i.e., $|\Phi| = 1$. Suppose the players' action and payoffs are as in Figure 5. Suppose payoffs are deterministic. This game has a unique Nash equilibrium (in pure and mixed strategies) (T, L). This Nash equilibrium is strict. Moreover, (T, L) is the Stackelberg outcome.

Instead of the statement "there exists a period K such that if $Q^K(s, a_Q) \in \mathcal{O}_{(s,a_Q)}$" Theorem 5 requires that $Q^0(s, a_Q) \in \mathcal{O}_{(s,a_Q)}$ as well as conditions on the sequences (α^t) and (ε^t). The hypothesis on the initial Q-values is required because the event $Q^K(s, a_Q) \in \mathcal{O}_{(s,a_Q)}$ need not be measurable with respect to the short-run players' information P^K. The requirements on (α^t) and (ε^t) can be weakened. More precisely, for any $(\tilde{\alpha}^t)$ and $(\tilde{\varepsilon}^t)$, there exists a $K \in \mathbb{N}$ such that the parameters can be chosen to satisfy $\alpha^t = \tilde{\alpha}^{t+K}$ and $\varepsilon^t = \tilde{\varepsilon}^{t+K}$.

[63] The hypotheses of Theorems 3 and 6 imply that this necessary condition is satisfied.

Claim 3. *Suppose the algorithm is transparent. There exists a $\xi > 0$ and an open set of parameters $\langle Q^0, (\alpha^t), (\varepsilon^t)\rangle$ for the algorithm such that*

$$\mathbb{P}\left[\{Q^t(T) > \max\{Q^t(M), Q^t(B)\} \quad \textit{for infinitely many } t\}\right] \leq 1 - \xi.$$

Proof in Appendix **??**

The claim states that, for some parameters of the algorithm, there is a positive probability that the algorithm does not learn the Stackelberg action. Consequently, play fails to converge to a Nash equilibrium of the auxiliary game.[64] ■

B.2. Sufficient conditions for learning the Stackelberg action

In this section, we present conditions that are sufficient for the algorithm to learn the Stackelberg action.

Theorem 6. *Let*

$$A_Q^* = \{a_Q \mid \exists s \in S_Q : a_Q^{\text{Stack}}(s) = a_Q\}.$$

Suppose for each action $a_Q \in A_Q^*$ there exists $s \in S_Q$ such that equation (1) *holds, i.e.,*

$$\min_{a_{\text{SR}}} \mathbb{E}[u_Q(a_Q, a_{\text{SR}}, \omega)|s] > \max_{a_{\text{SR}}} \mathbb{E}[u_Q(a_Q', a_{\text{SR}}, \omega)|s] \ \forall a_Q \neq a_Q'. \qquad (1)$$

When the signals about the algorithm's action are precise enough, in every equilibrium, the algorithm learns to play the Stackelberg action and receives approximately the Stackelberg payoff in each state.
Formally, there exists $\overline{\gamma} > 0$ such that for all γ-perfect monitoring structures (Φ, p) with $\gamma \leq \overline{\gamma}$, for all parameters $\langle Q^0, (\alpha^t), (\varepsilon^t)\rangle$ of the algorithm and for every state $s \in S_Q$,

$$\lim_{T\to\infty} \frac{\sum_{t=0}^{T} \mathbb{1}\left\{a_Q^t = a_Q^{\text{Stack}}(s)\right\} \mathbb{1}\left\{\omega^t \in s\right\}}{\sum_{t=0}^{T} \mathbb{1}\left\{\omega^t \in s\right\}} = 1,$$

and

$$\lim_{T\to\infty} \frac{\sum_{t=0}^{T} u_Q\left(a_Q^t, a_{\text{SR}}^t, \omega^t\right) \mathbb{1}\left\{\omega^t \in s\right\}}{\sum_{t=0}^{T} \mathbb{1}\left\{\omega^t \in s\right\}} \in \left((1-\gamma)u_Q^{\text{Stack}}(s) - \gamma M, (1-\gamma)u_Q^{\text{Stack}}(s) + \gamma M\right)$$

almost surely in every equilibrium. Here, M is a bound on the norm of the algorithmic player's expected payoff.
Proof in Appendix B.3.3.

[64] By virtually the same argument, one can show that play does not converge to the set of correlated equilibria. [65]

		SR	
		L	R
algorithmic player	T	$2, 1$	$4, 0$
	B	$3, 0$	$0, 1$

Figure 6: Payoff functions such that the algorithm achieves the Stackelberg payoff but the sufficient conditions of Theorem 6 fail.

Theorem 6 provides weaker sufficient conditions than 3 that ensures the algorithm learns the Stackelberg action in every state. The sufficient conditions are weaker in two respects.

First, the hypothesis that for each $a_Q \in A_Q^*$ equation (1) holds for some state $s \in S_Q$ is weaker than the hypotheses in Theorem 3. Assumption 3 and the hypothesis that the algorithm's state space is a rich partition according to Definition 6 together imply that for each a_Q equation (1) holds for some $s \in S_Q$.

Second, Theorem 6 makes weaker assumptions on the short-run players' information about the outcome of past interactions.[66] Theorem 3 assumed that the short-run players have no information about past outcomes. Theorem 3 allows for short-run players to observe all past outcomes or even a subset thereof. For example, Theorem 6 includes the case where short-run players observe the algorithm's past action or past states, but not its realized payoff. Moreover, the short-run player in period t can have more information about outcomes before period $t-1$ than SR^{t-1}. Nonetheless, both theorems require that SR^t does not observe the algorithm's information about the realization of ω^t. This cannot be weakened without imposing stronger conditions on the (expected) payoff functions than equation (1).

The conditions of Theorem 6 are sufficient but not necessary for the algorithm to learn the Stackelberg outcome. To see this, consider the game with payoffs given in Figure 6. The Stackelberg payoff is 2 and is attained by the algorithmic player playing T. Suppose the algorithm is transparent so that the short-run player can condition her action on the Q-values. Suppose the signal ϕ is pure noise. Eventually, $Q^t(T) \geq 2$. Consider a period t such that $Q^{t-1}(B) < 2 < Q^t(B)$. Let τ be the next period T is played. Then $Q_{\tau+1}(B) < Q_{\tau+1}(T)$ almost surely. Moreover, if the algorithm plays B in $t+1$, then $Q^{t+2}(B) < 2 \leq Q^{t+1}(T)$. Consequently, the algorithmic player receives a payoff of 0 in a period t only if it experimented in period $t-1$. Because $\varepsilon^t \to 0$, the algorithm's long-run average payoff is above 2 almost surely.

[66] How precise the signals must be, i.e., the cutoff $\overline{\gamma}$ in the statement of Theorem 6, can be chosen independent of the short-run players' information about past outcomes, i.e., the partitions $\{P_t\}$.

		SR	
		L	R
	T	$3, 1$	$1, 0$
algorithmic player	M	$2, 1$	$0, 0$
	B	$0, 0$	$0, 1$

s_1

	SR	
	L	R
T	$0, 1$	$0, 0$
M	$1, 1$	$1, 0$
B	$0, 0$	$0, 1$

s_2

	SR	
	L	R
T	$0, 1$	$0, 0$
M	$0, 1$	$0, 0$
B	$1, 0$	$1, 1$

s_3

Figure 7: The payoff functions for Example 6. In state s_1, T is a strictly dominant action for the algorithm. Moreover, the Stackelberg payoff in s_1 is attained when playing the strictly dominant action. In states s_2 and s_3, the actions M and B, respectively, satisfy equation (1).

Can the condition in (1) be weakened to

$$\mathbb{E}[u_Q(a_Q, a_{\text{SR}}, \omega)|s] > \mathbb{E}[u_Q(a'_Q, a_{\text{SR}}, \omega)|s] \ \forall a'_Q \neq a_Q, a_{\text{SR}} \in A_{\text{SR}}?$$

That is, is it sufficient for Theorem 6 that for each action a_Q there exists $s \in S_Q$ such that this action a_Q is strictly dominant action for payoffs $\mathbb{E}[u_Q(\cdot, a_{\text{SR}}, \omega)|s]$? The following Example 6 shows that this is not the case.

Example 6. Suppose $A_Q = \{T, M, B\}$ and $A_{\text{SR}} = \{L, R\}$. Suppose the state space of the algorithm contains three elements, $S_Q = \{s_1, s_2, s_3\}$. Suppose that expected payoffs are given by the payoff matrices in Figure 7. The Stackelberg actions are T, M and B in states s_1, s_2 and s_3 respectively. Moreover, the Stackelberg action in each state is strictly dominant. Note, however, that equation (1) is not satisfied for the action T in any state.

Assume all states occur with probability 1/3. Let the signalling structure be given by $\Phi = \{\phi_T, \phi_M, \phi_B\}$ and, for an $x \in (0, 1)$,

$$\begin{aligned} p(\phi_T|T) = p(\phi_M|M) = p(\phi_B|B) = 1 - \gamma; \\ p(\phi_M|B) = p(\phi_M|T) = p(\phi_T|B) = p(\phi_B|T) = \gamma/2; \\ p(\phi_T|M) = x\gamma, p(\phi_B|M) = 1 - \gamma - x\gamma. \end{aligned}$$

Let[67]

$$\begin{aligned} Q^0(s_1, T) = 1, Q^0(s_1, M) = 2, Q^0(s_1, B) = 0; \\ Q^0(s_2, M) = 1, Q^0(s_2, T) = Q^0(s_2, B) = 0; \\ Q^0(s_3, B) = 1, Q^0(s_3, T) = Q^0(s_3, T) = 0. \end{aligned}$$

[67] It suffices to assume that the Q-values are in a neighborhood of the ones specified here.

Then the algorithm plays according to

$$\tau_Q : s_1 \to M, s_2 \to M, s_3 \to B$$

in period 0, up to experimentation. For $x > 0$ and $\gamma > 0$ small enough, the unique best-response of the short-run player to τ_Q is

$$\tau_{\text{SR}} : \phi_T \mapsto R, \phi_M \mapsto L, \phi_B \mapsto R.$$

Given this strategy, the expected payoff to the algorithm in state s_1 when playing T is $1 + \gamma$ and when playing M is $2(1 - \gamma)$. The latter is strictly larger than the former for $\gamma < 1/2$. Consequently, $(\tau_Q, \tau_{\text{SR}})$ is a strict equilibrium of the auxiliary extensive-form game. By Theorem 5, there exists parameters $\langle Q^0, (\alpha^t), (\varepsilon^t)\rangle$ such that $\mathbb{E}[u_Q(a^t_Q, a^t_{\text{SR}}, \omega^t)|s_1] < 5/2$ for all t. Moreover, this holds irrespective of the short-run players' information about past play.

The example shows: for any $\overline{\gamma}$ there exists a γ-perfect signalling structure, $0 < \gamma < \overline{\gamma}$ and parameters of the algorithm such that the limit of the algorithm's expected average payoff is bounded away from the state-by-state Stackelberg payoff. ■

B.3. Proofs

B.3.1. Proof of Theorem 5

Proof. Fix $\xi > 0$.

Let $\eta > 0$ be such that for each $s \in S_Q$,

$$\begin{aligned}&\sum_{\phi\in\Phi} p(\phi|\tau_Q(s))\mathbb{E}[u_Q(\tau_Q(s), \tau_{\text{SR}}(\phi), \omega)|s] - \eta \\ &> \sum_{\phi\in\Phi} p(\phi|a_Q)\mathbb{E}[u_Q(a_Q, \tau_{\text{SR}}(\phi), \omega)|s] + \eta \ \forall a_Q \neq \tau_Q(s).\end{aligned}$$

Since S_Q is finite, such a $\eta > 0$ exists by the hypothesis that $(\tau_Q, \tau_{\text{SR}})$ is a strict Nash equilibrium of $G(S_Q, (\Phi, p))$.

Let $\mathcal{O}_{(s,a_Q)}$ be the open set

$$\left(\sum_{\phi\in\Phi} p(\phi|a_Q)\mathbb{E}[u_Q(a_Q, \tau_{\text{SR}}(\phi), \omega)|s] - \eta, \sum_{\phi\in\Phi} p(\phi|a_Q)\mathbb{E}[u_Q(a_Q, \tau_{\text{SR}}(\phi), \omega)|s] + \eta\right),$$

for all $s \in S_Q, a_Q \in A_Q$.

Let $(\tilde{\alpha}^t)$ and $(\tilde{\varepsilon}^t)$ be sequences of updating parameters and experimentation rates

satisfying Assumptions (Step-Size) and (Experimentation), respectively.
Because $(\tau_Q, \tau_{\text{SR}})$ is a strict Nash equilibrium of $G(S_Q, (\Phi, p))$, there exists $\xi_1 > 0$ such that τ_{SR} is the unique best-response when the algorithmic player plays $\tau_Q(s)$ with probability at least $1 - \xi_1$ for all $s \in S_Q$.

Fix a $0 < \xi_2 < \min\{\xi, \xi_1\}$ to be determined below. As $\tilde{\varepsilon}^t \to 0$, there exists $M_1 \in \mathbb{N}$ such that for all $t \geq M_1(\xi_2)$, $(1\varepsilon_t)(1 - \xi_2) \geq 1 - \xi_1$.

Hence, for any $t \geq M_1$ and $p^{t-1} \in P^{t-1}$, such that

$$\mathbb{P}\left[\forall s \in S_Q, Q^t(s, \tau_Q(s)) > Q^t(s, a_Q), a_Q \neq \tau_Q(s) \mid p^{t-1}\right] \geq 1 - \xi_2,$$

the short-run player SR^t's (unique) best-response is $\sigma_{\text{SR}^t}(p^{t-1}, \phi) = \tau_{\text{SR}}(\phi)$.

Consequently, it suffices to show that there are $\langle Q^0, (\alpha^t), (\varepsilon^t)\rangle$ such that, for all s, t,

$$Q^t(s, \tau_Q(s)) > Q^t(s, a_Q) \; \forall a_Q \neq \tau_Q(s) \tag{4}$$

with probability at least $1 - \xi_2$. For each (s, a_Q) choose $Q^0(s, a_Q) \in \mathcal{O}_{(s,a_Q)}$. By definition of $\mathcal{O}_{(s,a_Q)}$, equation (4) holds for $t = 0$. Because the payoffs are sub-Gaussian, we can apply Lemma 6 to each of the processes $Q^t(s, a_Q)$. Hence, there exists M_2 such that, if

$$Q^{M_2}(s, a_Q) \in O_{(s,a_Q)},$$

then the probability of the event

$$\left\{\forall s, a_Q, t \geq M_2, Q^t(s, a_Q) \in \mathcal{O}_{(s,a_Q)}\right\}$$

occurs with probability at least $1 - \xi_2$ if the algorithm's parameters are $\langle Q^0, (\tilde{\alpha}^t), (\tilde{\varepsilon}^t)\rangle$. Denote $M = \max\{M_1, M_2\}$. Choosing $\alpha^t = \tilde{\alpha}^{M+t}, \varepsilon^t = \tilde{\varepsilon}^{M+t}$ and Q^0 as before,

$$\left\{\forall s, a_Q, tQ^t(s, a_Q) \in \mathcal{O}_{(s,a_Q)}\right\}$$

occurs with probability at least $1 - \xi_2$. The claim follows. □

B.3.2. Proof of Claim 3

Proof. Choose $Q^0(T), Q^0(B) < 0$ and $Q^0(M) \in (1, 1 + \delta)$ for a $\delta > 0$. There exists a $\overline{\varepsilon} > 0$ such that if the algorithmic player plays M with probability at least $1 - \overline{v}$, the short-run player's unique best-response is M. Choose the sequence (ε^t) to satisfy $\overline{\varepsilon} > \varepsilon^t$.

Observe that, as long as all SR^t play M, $Q^t(T) \leq 0$. Note that, with probability 1, there exists τ such that $Q^\tau(B) < Q^\tau(M) < Q^{\tau+1}(B)$ and $Q^\tau(T) < Q^\tau(B)$. Note that this implies that the algorithm experiments in period $\tau + 1$; for otherwise, it selects the greedy action M and $Q^\tau(B) = Q^{\tau+1}(B)$. Note also that in period $\tau + 2$, the short-run player plays R.

Fix such a τ. If the algorithm does not experiment in periods $\tau+2,\dots,\tau+6$, $Q^{\tau+6}(M) > 1 > \max\{Q^{\tau+6}(B), Q^{\tau+6}(T)$. To see this, note that in these periods (B,R) is played unless the algorithm experiments. Expanding the Q-values, one sees that $Q^{\tau+6}(B) < 1$ when $Q^{\tau} < 1$ and the algorithm receives the payoff 4 in $\tau+1$, but a payoff of -0.01 in $\tau+2,\dots,\tau+6$.[68]

Consequently, $Q^t(T) > 0$ only if there exists a $k < t$ such that the algorithm experiments at least twice in periods $k,\dots,k+6$. It is easy to see that there exists a sequence (ε^t) that satisfies Assumption (Experimentation) such that

$$\mathbb{P}\left[\left\{\exists t \mid \sum_{j=0}^{6} \varepsilon^{t+j} \geq 2\right\}\right] < 1-\xi$$

for a $\xi > 0$. The claim follows. □

B.3.3. Proof of Theorem 6

Proof. The proof follows along the lines of the proof of Theorem 3. The main change is to use Lemma 4 instead of Lemma 1.

Details are omitted.

□

[68] Here, $\alpha^{t+1} \geq (1-\alpha^{t+1})(1-\alpha^t)$ is used.