# AI in Action: Algorithmic Learning with Strategic Consumers

Stephan Waizmann *†

October 26, 2024
Job Market Paper
Click here for latest version.

This paper investigates the impact of artificial intelligence on the interaction between firms and consumers. It focuses on the use of learning algorithms in environments with strategic consumers — where learning must occur in the face of consumers who best-respond and adapt their behavior. An algorithm is transparent if consumers observe its inputs, whereas it is opaque if consumers do not observe its inputs. The main results are as follows. First, opaque algorithms perform better for the firm than transparent ones. In contrast to a transparent algorithm, an opaque algorithm learns the optimal policy and maximizes long-run profits. Second, opaque algorithms outperform transparent ones in terms of consumer welfare in important applications. That is, consumers may benefit from having less information about the algorithm's inputs. Third, whether the firm benefits from using an algorithm instead of behaving strategically depends on consumers' information about the algorithm's inputs. When the algorithm is opaque, it yields higher payoffs than a fully strategic firm.

*Department of Economics, Yale University, stephan.waizmann@yale.edu.

† I am indebted to Johannes Hörner, Larry Samuelson, and Marina Halac for their invaluable support and guidance. I am grateful to Amirreza Ahmadzadeh, Dirk Bergemann, V. Bhaskar, Hector Chade, Joyee Deb, Daniel Hauser, Michelle Hyun, Doron Ravid, Bernardo Ribeiro, Anna Sanktjohanser, Philipp Strack, Roland Strausz, Caroline Thomas, Juuso Välimäki, Allen Vong, Kai Hao Yang, and audiences at the Yale Micro lunch and the ESIF Economics and AI+ML Meeting 2024 for helpful discussions. Part of this work was conducted during my stay at the Toulouse School of Economics, whose hospitality is greatly acknowledged. All errors are my own.

# 1. Introduction

The increasing integration of artificial intelligence (AI) in business operations marks a significant shift in how companies interact with consumers. From content recommendation to fraud detection to dynamic pricing, firms use AI in their interaction with consumers in various applications. We explore AI's effect on the interaction between firms and consumers, focusing on a core application of AI: learning algorithms.

Learning algorithms enable firms to detect patterns and adapt strategies based on predictive analytics. However, when a firm employs them in its interaction with consumers, learning must occur in the face of consumers that best-respond and therefore adapt their behavior. Despite their widespread use, little is known about how learning algorithms perform in such a strategic environment.

In this paper, we examine the interaction between a learning algorithm and strategic consumers. We investigate to what long-run outcomes the use of learning algorithms in firm-consumer relations leads — in terms of profits, consumer welfare, and conduct.

To address these questions, we propose a model of a learning algorithm interacting with consumers. In our model, a firm repeatedly interacts with consumers. The firm behaves according to a reinforcement learning algorithm. Consumers, on the other hand, best-respond to the firm's (expected) decisions. We place no restrictions on the nature of the interaction between the firm and consumer. The model thus nests applications as diverse as content recommendation, pricing, and quality provision.

We consider a standard class of learning algorithms that is frequently employed and studied. The purpose of these algorithms is to determine the optimal policy in the long-run through taking actions and observing realized profits. They work as follows. The algorithm uses the outcomes of past interactions with consumers — the actions it has taken and the profit that obtained — to form an estimate of each action's profitability. It then either chooses the action for which its estimated profitability is highest or it experiments with an arbitrary action. After profits realize, it updates its estimate.

As we show, the long-run dynamics depend critically on the information consumers have about the algorithm, specifically whether the algorithm is transparent or opaque. We call an algorithm transparent if consumers observe its inputs, i.e., the outcomes of past interactions of the algorithm with consumers. It is opaque if consumers do not observe its inputs.

Our main results are as follows. First, we show that opaque algorithms perform better for the firm than transparent algorithms. In contrast to transparent algorithms, opaque algorithms converge to the optimal policy even in this strategic environment. As a consequence, opaque algorithms yield higher profits than transparent algorithms. The algorithm thus benefits from consumers having less

information about its inputs. Intuitively, when consumers have less information about the algorithm's inputs, they are more reactive to information about the algorithm's current behavior. This in turn enables the algorithm to learn about the environment and so to play the optimal action.

Second, perhaps surprisingly, opaque algorithms not only raise firm's profits but may lead to higher consumer surplus as well. Indeed, for a large class of games, consumer welfare is higher when the algorithm is opaque than when the algorithm is transparent. That is, consumers, on average, can be better off when they have less information about the algorithm. While each individual consumer benefits from more information, her behavior affects the algorithm's learning and thus poses an externality on future consumers. If consumers benefit from the algorithm's learning, having less information about the algorithm creates a positive externality on other consumers. We provide conditions under which the externality is positive so that consumers, on average, gain when the algorithm is opaque instead of transparent. This is satisfied, for instance, if the algorithm recommends products to the consumer, or if the algorithm decides on the quality level of a service.

Third, the transparent-opaque dichotomy also matters when comparing the algorithm's performance to the profits the firm could obtain as a strategic player. Comparing the long-run profits of the algorithm to the ones a strategic player can obtain, the following holds: in the transparent case, the strategic player obtains higher payoffs than the algorithm; in the opaque case, the algorithm receives weakly higher payoffs than the strategic player. In both cases, the difference in payoffs is strict for some games.

From a methodological point of view, we ask which properties of reinforcement learning algorithms carry over to a strategic environment. In a non-strategic, stationary Markov environment, the algorithms we consider learn the policy that is optimal under complete information about the environment. Moreover, the algorithm asymptotically achieves this maximum payoff. We characterize conditions under which these two properties — learning the optimal policy and attaining the maximum payoff, subject to consumers playing a best-response — carry over to environments with short-lived consumers.[1]

We contribute to three distinct literatures. First, our paper adds to the understanding of how artificial intelligence impacts markets with a focus on learning in environments of incomplete information. Second, we contribute to the literature on reinforcement learning in multi-agent environment. The main difference to the literature is that we consider the interaction between one algorithm and a (myopically) best-responding agent. Third, we contribute to the literature on

[1]The environment with a single short-lived consumer in each period is a minimal departure from the exogenous, stationary environment — where learning algorithms are well understood — to a strategic environment.

learning in repeated games. Section 4 discusses the related literature in-depth.

The remainder of the paper is organized as follows. Section 2 introduces the model. In Section 3, we present the main results. Section 4 discusses our contribution to the literature in detail. Section 5 concludes. Proofs are relegated to the Appendix. A supplementary Online Appendix contains additional results and examples.[2]

# 2. Model and algorithm

## 2.1. Model

**Environment** Time is discrete and infinite, $t = 0, 1, \ldots$ There is one long–lived player, called the algorithm or algorithmic player, who is active in every period. In each period $t$, there is a short-run player $\text{SR}^t$ who is active only in that period. Let $A_Q$ and $A_{\text{SR}}$ be two finite set of actions of the algorithmic player and the short–run player $\text{SR}^t$, respectively.

We assume the environment changes over time, reflecting the notion that what constitutes optimal play may evolve. Let $(\omega^t)_t$ be a sequence random variables with support $\Omega$. For simplicity, assume $(\omega^t)$ are iid. Denote the distribution of the random variable $\omega^t$ by $q(\cdot)$, i.e., $\mathbb{P}[\omega^t = \omega] = q(\omega)$. Throughout, we assume that $\Omega$ is countable.

**Information** To allow players to react to changes in the environment, assume that in each period $t$, the algorithmic player and the short–run player $\text{SR}^t$ receive some information about the realization of $\omega^t$. We model this information via partitions. Let $S_Q$ and $S_{\text{SR}}$ be two finite partitions of $\Omega$. After $\omega^t$ has realized, the algorithmic player is informed of the cell $s \in S_Q$ that contains $\omega^t$, i.e., the unique $s \in S_Q$ such that $\omega^t \in s$. Likewise, the short–run player $\text{SR}^t$ observes the cell in $s' \in S_{\text{SR}}$ that contains $\omega^t$.

Throughout, we assume that the algorithmic player has more information about the realized random variable than the short–run players do.

**Assumption 1.** *The partition $S_Q$ is (weakly) finer than $S_{\text{SR}}$.*

We refer to the cell $s \in S_Q$ that contains the realized $\omega^t$ as the *state of the world* in period $t$.

The random variable $\omega^t$ and information about it matter for payoffs. Given a realization $\omega \in \Omega$, we are given a pair of payoff functions that map joint actions

[2]The Online Appendix is available at https://stephanwaizmann.github.io/website-docs/jmp_online_appendix.pdf.

into real numbers:[3]

$$u_Q(\cdot,\cdot,\omega), u_{SR}(\cdot,\cdot,\omega) : A_Q \times A_{\text{SR}} \to \mathbb{R}.$$

Furthermore, we allow for imperfect monitoring of actions. To this end, define a signalling structure $(\Phi, p)$ as follows: $\Phi$ is a finite set of signals and, for each $a_Q \in A_Q$, $p(\cdot|a_Q) \in \Delta(\Phi)$ is a probability distribution over signals. The signal $\phi$ is drawn independently from $\omega$, conditional on $a_Q$, and privately observed by the short-run players.

The following example showcases the components of the model.

**Example 1.** Consider a long-lived retailer that repeatedly interacts with a sequence of consumers. In each period, the consumer owns a product that is defective and has to decide if she asks for the product to be repaired or if she returns the product. The retailer decides whether to provide customer service in–house or to outsource customer service. In–house customer service provides a higher quality for the consumer. Outsourced customer service, however, is of lower quality.

Payoffs are as in Figure 1. The consumer prefers a repair if the quality of the customer service is high. If the customer service has a low quality, the consumer prefers to return the product. The service quality affects the consumer's payoff both when she returns the product and when she asks for a repair.

To allow for the possibility that the retailer's optimal policy changes over time, we assume its profit depends on the state of the economy. In each period, the economy can be in one of three states $\omega_i, i = 1, 2, 3$. The retailer observes the realized state at the start of each period. Formally, this means that $S_Q = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}\}$. For simplicity, denote by $s_i = \{\omega_i\}$ the state that contains the realized shock $\omega_i$. We return to this example below. ■

**Timing** Play evolves as follows. In period $t$, $\omega^t$ is drawn according to $q$. The algorithmic player and the short–run player $\text{SR}^t$ observe the cell of their information partition that contains $\omega^t$, i.e., $s_Q^t \ni \omega^t, s_{\text{SR}}^t \ni \omega^t$. Then the algorithmic player chooses an action $a_Q \in A_Q$. Next, a signal $\phi \in \Phi$ is drawn according to $p(\cdot|a_Q)$. $\text{SR}^t$ observes the realized signal and chooses an action $a_{SR} \in A_{SR}$. Play moves to

[3]Information partitions allow us to model random payoffs as well. For $a \in A_Q \times A_{\text{SR}}$ and $s \in S_Q$, denote $v_Q(a, s) = \mathbb{E}[u_Q(a, \omega)|\omega \in s]$. Then the payoff the algorithmic player receives when its information is $s$ is the random variable

$$u_Q(a, \omega)_{|\omega \in s} = v_Q(a, s) + \eta_{a,s}(\omega),$$

where $\eta_{a,s}(\omega) = u_Q(a, \omega)_{|\omega \in s} - \mathbb{E}[u_Q(a, \omega)|\omega \in s]$ is a payoff shock with mean zero. Information partitions are flexible enough to capture arbitrary correlations between the payoff shocks $\eta_{a,s}(\omega)$ for different action profiles $a$.

| | | Consumer | | Consumer | | Consumer | |
|---|---|---|---|---|---|---|---|
| | | repair | return | repair | return | repair | return |
| Retailer | high quality | $2,3$ | $0,2$ | $2,3$ | $0,2$ | $2,3$ | $0,2$ |
| | low quality | $-1,0$ | $-2,1$ | $3,0$ | $1,1$ | $4,0$ | $3,1$ |
| | | $\omega_1$ | | $\omega_2$ | | $\omega_3$ | |

Figure 1: The payoff functions for the Example 1.

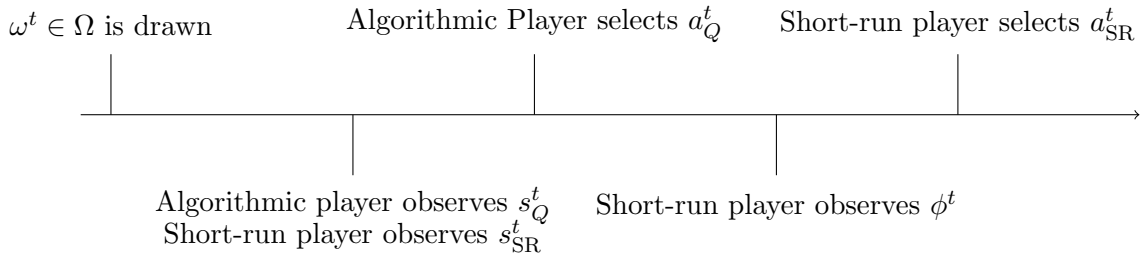


Figure 2: Timing

$t+1$. We describe what the players observe about the outcome of this and past interactions below. Figure 2.1 describes the timing of the interaction.

**Algorithmic approach** The main premise of this paper is that the algorithmic player does not choose its strategy. Instead, it plays according to a fixed $Q$–learning algorithm.

The algorithm is model-free: it does not depend on the primitives of the model (i.e., the payoff functions and probability distribution of the random variable $\omega^t$). The algorithm takes as given its state space $S_Q$ and its actions $A_Q$. All it takes as inputs are states and own past realized payoffs in those states, given the action it has taken.

$Q$–learning is a popular reinforcement learning algorithm and serves as a foundational model for many variations and enhancements applied by firms. Moreover, $Q$-learning has the advantage that it is relatively tractable. Therefore, it has become the workhorse model in the economics literature.[4] We view $Q$-learning as a proxy for more sophisticated reinforcement learning algorithms.

We expand on the definition of $Q$–learning in Section 2.2. In short, such an algorithm is described by a vector $\langle Q^0, S_Q, (\alpha^t), (\varepsilon^t)\rangle$. $S_Q$ is the algorithm's information about $\omega$. $Q^0 : S_Q \times A_Q \to \mathbb{R}$, and $Q^0(s, a_Q)$ is the initial guess about its payoffs from playing $a_Q \in A_Q$ given information $s \in S_Q$. The initial guess $Q^0$ captures all the information about the payoffs that is coded into the algorithm, i.e.,

[4]See the discussion in Section 4.

all the training of the algorithm that has been done before the interaction with the short–run players. The sequence $(\alpha^t)$ are updating parameters and $(\varepsilon^t)$ are experimentation probabilities. The standard assumptions on those are introduced below (see Assumptions (Step-Size) and (Experimentation) in Section 2.2). The parameters $\langle Q^0, S_Q, (\alpha^t), (\varepsilon^t)\rangle$ are common knowledge.

While the algorithmic player's strategy is exogenously fixed, the short-run players choose their actions strategically. This is where our paper departs from the literature on algorithmic learning; because each short-run player best replies, the environment is not exogenous. The environment with a single short-run player in each period is a minimal departure from the exogenous, stationary environment — where learning algorithms are well understood — to a strategic environment.

**Transparent and opaque algorithm** Lastly, we describe what the short-run players observe about the interaction of the algorithm with previous short-run players. To elucidate how the short-run players' information impacts the algorithm's learning, we focus on two extreme cases:[5] *transparent* algorithms and *opaque* algorithms. When the algorithm is transparent, the short-run player has the same information the algorithm has about past play and the current realization of the random variable $\omega$. In contrast, when the algorithm is opaque, the short-run players have no information about past interactions and no information about the algorithm's state.

When the algorithm is transparent, the short-run player in period $t$ observes the past interactions of the algorithm and the previous short-run players. That is, $\text{SR}^t$ observes the state $s_Q^k \in S_Q$, the algorithm's action $a_Q^k \in A_Q$, and the algorithm's realized payoff $u_Q^k$ in all periods $k$ prior to $t$.[6] Moreover, the short-run player $\text{SR}^t$ observes the state of the world $s_Q^t$ in period $t$. However, $\text{SR}^t$ does not observe the action the algorithm takes in period $t$, $a_Q^t$, but only the private signal $\phi^t \in \Phi$ about the action.

Denote by $U_Q$ the range of the algorithm's payoffs, i.e.,

$$U_Q = \{u_Q(a,\omega) | a \in A_Q \times A_{\text{SR}}, \omega \in \Omega\}.$$

**Definition 1** (Transparent Algorithm)**.** *An algorithm is* transparent *if the short-run players observe the algorithm's inputs. Formally, a history $h^t$ for player $\text{SR}^t$ is an element of $(S_Q \times A_Q \times U_Q)^{t-1} \times S_Q \times \Phi$. In that case, a strategy for player*

[5]These cases are not exhaustive. For example, the short-run players may observe the algorithm's past actions and realized payoffs, but not its state. The Online Appendix provides results for such intermediate cases.

[6]The results do not change if the short-run player $\text{SR}^t$ observes the actions $a_{\text{SR}}^k$ of the short-run player or the signal $\phi^k$ from previous periods.

$\mathrm{SR}^t$ *is a map from the set of histories into actions,*

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

In contrast, an algorithm is opaque if the short-run players have no information about its past interactions nor about its current information about $\omega$. Formally, the short-run players' information partition is trivial, $S_{\mathrm{SR}} = \{\Omega\}$. When the algorithm is opaque, the short-run player in period $t$ can condition her action only on the realized signal $\phi$ about the algorithm's action.[7]

**Definition 2** (Opaque algorithm)**.** *An algorithm is* opaque *if the short-run players do not observe its inputs. Formally, a strategy for* $\mathrm{SR}^t$ *is a map*[8]

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

Any strategy tuple of the short-run players $\sigma_{\mathrm{SR}} = (\sigma_{\mathrm{SR}^t})_t$ together with the algorithm induce a probability measure over outcomes $o \in (\Omega \times A_Q \times A_{\mathrm{SR}})$. Denote this probability measure and the expectation it induces by $\mathbb{P}^{\sigma_{\mathrm{SR}}}$ and $\mathbb{E}^{\sigma_{\mathrm{SR}}}$, respectively.[9]

The short-run players are Bayesian agents, and use the prior $q$ (as well as their knowledge of the signalling structure, etc.) to form expectations. A strategy tuple $\sigma^*_{\mathrm{SR}}$ is an equilibrium if for all players $\mathrm{SR}^t$

$$\mathbb{E}^{\sigma^*_{\mathrm{SR}}}\left[u_{\mathrm{SR}}(a^t_Q, a^t_{\mathrm{SR}}, \omega^t)\right] \geq \mathbb{E}^{(\sigma^*_{\mathrm{SR}^k})_{k\neq t}, \sigma'_{\mathrm{SR}^t}}\left[u_{\mathrm{SR}}(a^t_Q, a^t_{\mathrm{SR}}, \omega^t)\right] \ \forall \sigma'_{\mathrm{SR}^t}.$$

Throughout, we focus on strategies $\sigma_{\mathrm{SR}}$ that constitute an equilibrium. All results hold in any equilibrium.[10]

**Additional assumptions** We impose the following assumptions on the primitives of the model. In order to focus on the algorithm's learning, we abstract away from signalling considerations of the algorithmic player. We therefore impose a "known-own-payoffs"-assumption.

**Assumption 2** (Known-own payoffs)**.** *For all* $s_{\mathrm{SR}} \in S_{\mathrm{SR}}, \omega, \omega' \in s_{\mathrm{SR}}, a \in A_Q \times A_{\mathrm{SR}}$*,*

$$u_{\mathrm{SR}}(a, \omega) = u_{\mathrm{SR}}(a, \omega').$$

[7] The short–run player $\mathrm{SR}^t$ knows the period $t$ in which she is active.

[8] We suppress the dependence of $\sigma_{\mathrm{SR}^t}$ on $S_{\mathrm{SR}}$ because the partition is assumed to be trivial.

[9] We suppress dependence of the probability and the expectation operator on $\sigma_{\mathrm{SR}}$ when there is no chance of confusion.

[10] Our assumptions guarantee that the equilibrium is generically unique.

Assumption 2 requires that the short–run players have enough information to determine their own payoffs. In particular, even when the algorithm has more information about $\omega$ than the short–run players, the algorithm has no more information about the short-run players' payoff than they have. This implies that the algorithm's action does not convey information about the short-run players' payoff other than the action itself.[11]

In addition, we suppose that any preference ordering of the algorithmic player is possible.

**Assumption 3** (Richness Condition)**.** *For any strict preference relation $\succ$ on $A_Q \times A_{\text{SR}}$ there exists $\omega \in \Omega$ such that $u_Q(\cdot, \omega)$ represents $\succ$, i.e., $a \succ a' \iff u_Q(a, \omega) > u_Q(a', \omega)$.*

Assumption 3 requires that for all preference relations over joint action pairs, there exists a realization of the random variable $\omega$ such that the corresponding payoff function of the algorithmic player represents this preference relation. Informally, Assumption 3 requires that the set of possible payoff functions of the algorithmic player is *rich*. Interpreting the random variable $\omega$ as payoff shocks, Assumption 3 is satisfied if the support of the payoff shocks is large. In particular, Assumption 3 is satisfied if payoff shocks are independent across actions and have unbounded support.[12] We maintain Assumptions 2 and 3 throughout.

Throughout, we make two further technical assumptions. Unless explicitly stated, these assumptions are maintained.

For each joint action $a \in A_Q \times A_{\text{SR}}$ and state $s \in S_Q$, the conditional distribution of payoffs is sub-Gaussian;[13] that is, $u_Q(a, \omega) \sim q_{|s}$ is sub-Gaussian. Sub-Gaussian random variables have a finite variance. We remark that each distribution with a bounded support is sub-Gaussian.

Secondly, we assume that payoffs are generic. For all $s \in S_Q, a_Q \in A_Q$,

$$\mathbb{E}[u_{\text{SR}}(a_Q, a_{\text{SR}}, \omega)|s] = \mathbb{E}[u_{\text{SR}}(a_Q, a'_{\text{SR}}, \omega)|s] \iff a_{\text{SR}} = a'_{\text{SR}}.$$

[11] Under Assumption 1, Assumption 2 can be weakened to: $\forall s_{\text{SR}} \in S_{\text{SR}}, a \in A_Q \times A_{\text{SR}}$,

$$\mathbb{E}[u_{\text{SR}}(a, \omega)|s_Q] = \mathbb{E}[u_{\text{SR}}(a, \omega)|s_{\text{SR}}] \ \forall s_Q \subset s_{\text{SR}}.$$

[12] Our leading Example 1, as stated, does not satisfy Assumption 3. However, Assumption 3 is satisfied if the payoffs in Figure 1 are interpreted as expected payoffs in each state $s_i$.

[13] Recall that a random variable $\omega$ is sub-Gaussian if there exists a real number $r \in (0, \infty)$ such that $\mathbb{E}[\exp(\lambda\omega)] \leq \exp(\lambda^2 r^2/2)$ for each real number $\lambda$.

For all $s \in S_Q$ and $a \in A_Q \times A_{\text{SR}}$,

$$\mathbb{E}[u_Q(a,\omega)|s] = \mathbb{E}[u_Q(a',\omega')|s] \iff a = a'.$$

Genericity of the short-run player's payoff implies that the short–run players' best-response is unique if the algorithmic player takes an action with probability 1. If the short-run players' payoffs are not generic, there is little hope for the algorithm to converge, as the short-run players can alternate among multiple best-responses, thus preventing learning.

## 2.2. Q–learning

We provide a brief description of $Q$–learning. See Watkins (1989), Watkins and Dayan (1992) or, e.g., chapter 6.5 in Sutton and Barto (2018) for a detailed exposition. Readers familiar with $Q$–learning may wish to read the section to acquaint themselves with the notation used subsequently.[14]

Consider a single-player decision problem with state space $S_Q$, available actions $A_Q$, and (random) rewards $u(\cdot,\omega)$ where $\omega$ is distributed according to $q(\omega|s)$. Here, $q(\omega|s) = q(\omega)/\sum_{\omega' \in s} q(\omega')$ if $\omega \in s$ and $q(\omega|s) = 0$ otherwise.

$Q$-learning is a method to find the optimal policy in this decision problem. $Q$–learning is *model–free* in that it does not depend on the primitives $u(\cdot,\cdot)$ and $q(\cdot|\cdot)$; that is, $Q$-learning is designed to find the optimal policy when the rewards $u(\cdot,\cdot)$ and the probability distribution $q(\cdot|\cdot)$ are unknown.

Let the (unknown) value function of the decision problem be $V^*$. A useful concept is the *state–action value function* defined as

$$Q^*(s,a) = \mathbb{E}[u(a,\omega)|s].$$

By Bellman's Principle of Optimality, $\max_a Q^*(s,a) = V^*(s)$.

The aim of $Q$–learning is to find $Q^*$. The algorithm is designed as follows. Fix an initial guess $Q^0 : S_Q \times A_Q \to \mathbb{R}$. At period $t$, when the state is $s$ and the action $a$ is selected, the updated guess is

$$Q^t(s,a) = (1-\alpha^{n_t})Q^{t-1}(s,a) + \alpha^{n_t}u,$$

where $u$ is the realized payoff, $\alpha^t$ is an updating parameter, and $n_t$ is the number of visits to $(s,a)$ before period $t$. For all other state-action pairs $(\tilde{s},\tilde{a}) \neq (s,a)$, the guess is not updated, i.e.,

$$Q^t(\tilde{s},\tilde{a}) = Q^{t-1}(\tilde{s},\tilde{a}).$$

[14] Our description of $Q$–learning differs from standard descriptions since the changes in the environment, $\omega^t$, are assumed to be iid.

Assume that the algorithm chooses actions $\varepsilon^t$–greedily, i.e., according to the following rule. At period $t$ and state $s$, it chooses the "greedy" action

$$a \in \arg\max_{a'} Q^{t-1}(s, a')$$

with probability $1 - \varepsilon^t$, and with probability $\varepsilon^t$, each $a \in A_Q$ is chosen with probability $1/|A_Q|$.[15] Specifying the updating parameters $(\alpha^t)$ and the experimentation rates $(\varepsilon^t)$, the algorithm is well-defined.

Throughout, we make two assumptions on the parameters of the $Q$–learning algorithm, one for the learning rates and one for the experimentation probabilities.

**Assumption (Step-Size).** *The updating parameters* $(\alpha^t)$ *satisfy*

$$\sum_{t=0}^{\infty} \alpha^t = \infty$$

*and*

$$\sum_{t=0}^{\infty} \left(\alpha^t\right)^2 < \infty,$$

*as well as* $\alpha^{t+1} \geq \alpha^t(1 - \alpha^{t+1})$ *for all* $t$.

This assumption, also known as the Robbins and Monro (1951)-step-size condition, is standard in the literature on learning and stochastic approximation. Intuitively, the first part requires that each observation carries sufficient weight so that the impact of the initial guess and any single observation washes out. The second part requires that weights decay fast enough so that observing an outlier does not move the estimate too much. The last condition says that the weight attached to the payoff of the $t+1$-th update is at least as large as the weight attached to the payoff in the $t$-th update.[16]

**Assumption (Experimentation).** *The experimentation rates* $(\varepsilon^t)$ *satisfy*

$$\sum_{t=0}^{\infty} \varepsilon^t = \infty,$$

*and* $\varepsilon^t \to 0$ *monotonically as* $t \to \infty$.

[15] Assume ties are broken with equal probability.

[16] The condition $\alpha^{t+1} \geq \alpha^t(1 - \alpha^{t+1})$ is satisfied by commonly used specifications for the updating parameters, including $\alpha^t = \alpha/t^s, \alpha > 0, s \in (1/2, 1]$. While this condition is not needed for convergence, it simplifies the analysis.

The first condition says that the algorithm experiments often enough. The second condition guarantees that the algorithm eventually plays according to the optimal policy after it has learned it. Both conditions are needed to ensure that the algorithm learns the optimal policy and plays asymptotically according to it. Taken together, Assumptions (Step-Size) and (Experimentation) are the weakest set of assumptions in the literature that guarantees that the algorithm's asymptotic behavior is optimal in a single-player environment, i.e., an environment in which only the algorithm takes decisions. Thus, these assumptions are a starting point for our analysis.

**Theorem (Watkins).** *Assume that* $q(s) = \sum_{\omega \in s} q(\omega) > 0$ *for all states* $s \in S_Q$. *Then, under Assumptions (Step-Size) and (Experimentation),* $Q^t(s,a) \to Q^*(s,a)$ *almost surely as* $t \to \infty$ *for all state-action pairs* $(s,a) \in S_Q \times A_Q$.[17]

**Example 1** (continued)**.** We provide an alternative description of an algorithm. While the precise specifications differ from the $Q$-learning algorithms we consider in the sequel, readers who are not interested in technical details may wish to think of the algorithm as described here. Most of the intuition remains the same.

Suppose the algorithm keeps track of the average payoff received in state $s \in S_Q$ when having played the action $a_Q$. When state $s$ occurs, the algorithm either chooses the action that has generated the highest average payoffs in that state $s$ or it experiments and picks an action at random. The probability of experimentation decays to 0 but does so sufficiently slowly.■

# 3. Results

This section presents our main results. First, we discuss a benchmark, then present the results for transparent and opaque algorithms. We compare the outcomes for a transparent and for an opaque algorithm. Lastly, we compare the outcomes to a benchmark in which the long-run player is strategic. Throughout, we first illustrate the results using Example 1.

## 3.1. Benchmark

In this section, we present a benchmark in which the short-run players observe the algorithm's action perfectly before choosing their own actions. We argue that this implies that the environment is stationary in the following sense: in each period, the algorithmic player's payoff depends only on the action it takes in that period, and not on the actions taken or payoffs received in previous periods. We show that

[17] See Watkins and Dayan (1992) or Tsitsiklis (1994) for a proof.

the algorithm attains the maximal payoff, given that the short-run players play a best-response. This result obtains both for transparent and opaque algorithms.[18]

**Example 1** (continued)**.** Consider a benchmark case: before choosing whether to return the product or ask for a repair, the consumer observes the quality of the customer service. The optimal strategy of the consumer is immediate: ask for a repair if the service quality is high and return the product otherwise. Hence, the consumer's action does not depend on the history of play or on her information about the state. Taking this strategy as given, the profits the retailer receives are as follows: in state $s_1$, profits are higher when providing a high-quality service than when providing a low-quality service; in state $s_2$, profits are 2 when providing high quality but 1 when providing low quality; in state $s_3$, profits are strictly lower when providing a high-quality service than when providing low quality.

The algorithm then behaves as follows. Irrespective of the initial guess, the $Q$-values converge to the true profits: $Q^t(s_i, \text{low quality})$ converges to the profit the retailer receives when providing a low quality service and the consumer returns the product in every state $s_i$; similarly, $Q^t(s_i, \text{high quality})$ converges to the payoff the retailer receives when providing high quality and the consumer asks for a repair. The algorithm behaves asymptotically as follows: it provides high-quality service in states $s_1$ and $s_2$, but provides a low-quality service in state $s_3$. Hence, the algorithm takes the same action a fully informed retailer would have taken if it could commit to its preferred action. In the long–run, the retailer obtains the highest feasible complete information payoff, given that the consumer's decision is a best-response. ■

In this part, we assume that the short-run players observe the algorithm's action. Formally, the signalling structure is perfect.

**Definition 3** (Perfect signalling)**.** *Signalling is perfect if for each action $a_Q \in A_Q$ there exists a signal $\phi_{a_Q}$ such that $p(\phi_{a_Q}|a_Q) = 1$, and $|\Phi| = |A_Q|$.*

A posted-price mechanism is an example of a perfect signalling structure. The algorithm posts a price. The consumer observes the price and then makes her purchase decision.

Recall that, by Assumption 1, the algorithm has at least as much information about $\omega^t$ as the short-run player does. Hence, for any element $s \in S_Q$ there exists $s' \in S_{\text{SR}}$ such that $s \subset s'$.[19] For a cell $s' \in S_{\text{SR}}$ and an action $a_Q$ by the algorithmic

[18] A second natural benchmark is a setting in which the long-run player is not restricted to use a $Q$-learning algorithm. Instead, the long-run player can employ history-dependent strategies. We discuss this benchmark in Section 3.5.

[19] We use $\subset$ to denote weak set inclusion; in particular, $s \subset s'$ allows for the case $s = s'$.

player, the short-run player's best response is

$$\mathrm{BR}(a_Q, s') = \arg\max_{a_{\mathrm{SR}}} \mathbb{E}[u_{\mathrm{SR}}(a_Q, a_{\mathrm{SR}}, \omega)|s'].$$

By genericity, the best-response is unique. Assumption 2 implies that the short-run player's best-response does not change if given more information about the realized $\omega$: $\mathrm{BR}(a_Q, s') = \mathrm{BR}(a_Q, s)$ for any action $a_Q$ and any $s \subset s'$. Therefore, we abuse notation and denote the short-run player's best-response as a function of the algorithm's information $s \in S_Q$, even when the short-run player does not have the information $s$.

**Definition 4.** *The algorithmic player's* Stackelberg payoff in state $s \in S_Q$ *is*

$$u_Q^{\mathrm{Stack}}(s) = \max_{a_Q} \mathbb{E}[u_Q(a_Q, \mathrm{BR}(a_Q, s), \omega)|s].$$

Call the action of the algorithmic player that achieves the Stackelberg payoff in state $s$ the *Stackelberg action in state* $s$, denoted by $a_Q^{\mathrm{Stack}}(s)$. By genericity, the Stackelberg action is unique.

The Stackelberg payoff is the highest payoff the algorithmic player could achieve under two conditions. First, the algorithmic player has complete information about the payoff functions, including the short-run player's payoff function. Second, the algorithmic player can commit to play a pure action to which the short-run player best-responds.[20] Our first result states that perfect signalling is sufficient for the algorithm to learn the Stackelberg action.

**Theorem 1** (Benchmark: observed actions)**.** *Assume signalling is perfect. Then the algorithm learns to play the Stackelberg action and receives the Stackelberg payoff in each state.*

*Formally, for all parameters* $\langle Q^0, (\alpha^t), (\varepsilon^t)\rangle$*, and any state* $s \in S_Q$*,*

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

*and*

$$\lim_{T\to\infty} \frac{\sum_{t=0}^{T} u_Q(a_Q^t, a_{\mathrm{SR}}^t, \omega^t)\mathbb{1}\{\omega^t \in s\}}{\sum_{t=0}^{T} \mathbb{1}\{\omega^t \in s\}} = u_Q^{\mathrm{Stack}}(s)$$

*almost surely.*

*Proof in Appendix A.1.*

[20] By committing to a mixed action, the algorithmic player can achieve a weakly higher payoff than the pure Stackelberg payoff. Since a $Q$-learning algorithm does not play a mixed strategy asymptotically for generic parameters, the relevant comparison is to the pure Stackelberg payoff.

The intuition behind Theorem 1 is as follows. When the signal perfectly reveals the action of the algorithm, the action chosen by the short-run player depends only on the action played by the algorithm and her own information about the state. In particular, the short-run player's action is independent of the history (and the parameters of the algorithm). As a consequence, the payoff the algorithm receives depends only on the action it has taken and its own information $s$: the situation thus reduces to a single-player problem with random payoff

$$(s, a_Q, \omega) \mapsto u_Q\left(a_Q, \mathrm{BR}(a_Q, s), \omega\right).$$

The usual convergence result for $Q$-learning, Watkins's Theorem, then applies.

The first statement of Theorem 1 concerns the actions the algorithm plays. It states that the fraction of periods in which the algorithm takes the Stackelberg action converges to 1 almost surely. The action taken by the algorithm, however, does not converge almost surely because the algorithm experiments infinitely often with probability 1.[21]

Moreover, Theorem 1 states that the average payoff the algorithmic player receives in state $s$ converges almost surely to the Stackelberg payoff in state $s$. This is an asymptotic result; little can be said about the actions played by the algorithm and thus its payoff in the initial periods. Instead of considering the limit of average payoffs, we could consider the $\delta$-discounted expected payoff of the algorithm. Then the corresponding result would be that the expected discounted payoff converges to the Stackelberg payoff as the discount factor approaches 1.

We remark that Theorem 1 holds for transparent and opaque algorithms. When the short-run player $\mathrm{SR}^t$ observes the algorithm's action perfectly, her best-response does not depend on her belief about past play and the current state. The short-run player's action depends on the history only through the current, perfectly revealing signal. Hence, the result obtains irrespective of the information the short-run players have about the inputs of the algorithm.

We conclude that the convergence guarantees of $Q$–learning extend from single-player environments to environments in which the algorithm's actions are perfectly observed. The algorithm learns the optimal action and plays this action in almost every period asymptotically. Thus, the algorithmic player receives the same payoff it could have gotten (i) had it known the payoff functions of the game, including the payoff function of the short–run player, and (ii) could commit to playing an action. Perfect observability of its action guarantees that the $Q$-learning algorithm performs well. As we see next, this result relies on the signalling structure being perfect.

[21] In fact, the probability that, in state $s$, the algorithm takes the Stackelberg action converges to 1 almost surely. This is a stronger notion than the convergence of the empirical frequencies.

## 3.2. Transparent algorithm

In the following, we drop the benchmark assumption that the short-run players observe the algorithm's action perfectly. This implies that the environment in which the algorithm operates is non-stationary: the short-run players' behavior depends on their beliefs about the algorithm's inputs and thus on the outcome of past interactions. As we show, the outcomes differ depending on whether the algorithm is transparent or opaque.

In this section, we discuss the main results when the algorithm is transparent, i.e., when the short-run players observe its inputs. We show that the $Q$-learning algorithm typically fails to learn the Stackelberg outcome and to obtain the Stackelberg payoff. Hence, the two properties of $Q$-learning in single-player environments — learning the optimal action and attaining the maximal payoff — fail in the presence of strategic short-run players if the algorithm is transparent.

**Example 1** (continued)**.** Drop the assumption that the consumer observes the service quality before making her decision. Consider the case when the algorithm is transparent.

In states $s_1$ and $s_3$ learning obtains irrespective of the consumer's behavior: in state $s_1$, the retailer's profits when providing high quality are at least 0, so that the average realized payoff in state $s_1$ when providing high quality is eventually at least 0; similarly, the profits when providing low quality are at most $-1$ so that the average realized payoff for low quality in state $s_1$ eventually becomes negative. Consequently, the algorithm learns to play the action the retailer would choose to commit to under full information about the payoffs.

The situation is different in state $s_2$. For the sake of exposition, suppose the consumer's signal about the service quality is pure noise. When the algorithm is transparent, the consumer observes the outcome of past interactions. This enables the consumer to compute the average realized profit for each state-action combination in any period. Moreover, when the algorithm is transparent, the consumer knows the state. This implies that the consumer can predict the service quality the retailer provides, up to experimentation. Suppose in some period $t$, the average payoff when providing low quality in state $s_2$ is approximately 1, but the average payoff in the periods in which the algorithm provided high quality is strictly lower. Then the consumer expects the algorithm to provide low quality and hence chooses to return the product. Consequently, the payoff the retailer receives in state $s_2$ is 0 when experimenting and providing high–quality service but 1 when providing low quality. The average payoffs for low and high quality converge to 1 and 0, respectively. The algorithm then provides low quality in state $s_2$ in the long-run; the algorithm in state $s_2$ gets stuck in the Nash equilibrium of a simultaneous-move game with the same payoff functions. In particular, the

retailer does not achieve the profit it would have gotten if it had known the payoff functions and could commit to a quality level.

What happens if the consumer observes the service quality almost perfectly before making her decision? The consumer can then base her decision on the signal she observes. The consumer's posterior beliefs depend on the signal quality and her belief about the retailer's action. By design of the algorithm, the consumer expects the retailer to take the action that has yielded the highest average payoff so far, up to experimentation. As long as experimentation rates are large enough, the consumer's action is responsive to the signal she receives. However, as experimentation rates vanish, the consumer expects the retailer to take the action that has yielded the highest average payoff with probability close to 1. Hence, if the consumer expects the retailer to provide a low–quality service, she attributes an observation of any other signal to noise. As a consequence, the consumer's behavior becomes unresponsive to the signals: the situation becomes akin to the one in which the signal is pure noise. Since the experimentation probabilities vanish, there comes a time after which the consumer disregards the signals. ■

We dispose of the benchmark assumption that the short-run players observe the algorithm's action. Instead, we assume the signals about the algorithm's actions are noisy. Formally, we require that the signalling structure has full support.

**Definition 5** (Full–support signalling)**.** *The signalling structure* $(\Phi, p)$ *has* full support *if for all actions* $a_Q \in A_Q$*, each signal* $\phi \in \Phi$ *realizes with positive probability,* $p(\phi|a_Q) > 0$*.*

Signalling structures that have full support nest many interesting cases. Signalling has full support, for example, if a consumer has incomplete information about the quality of a good or service before making her purchase decision (as in Example 1). Moreover, full–support signalling includes the case when the signals are completely uninformative. This captures situations in which the algorithmic player and the short–run player move simultaneously.[22]

Recall that, when the algorithm is transparent, the short-run players and the algorithm have the same information about the state. That is, the information partitions of the short–run players and the algorithm coincide, $S_{\text{SR}} = S_Q$.

Fix a state $s \in S_Q = S_{\text{SR}}$. Define an auxiliary two-player simultaneous move game by $G(s) = \langle A_Q, A_{\text{SR}}, \mathbb{E}[u_Q(\cdot, \omega)|s], \mathbb{E}[u_{\text{SR}}(\cdot, \omega)|s]\rangle$. In this auxiliary game, the action sets correspond to the algorithmic player's and the short-run player's action sets, and the payoff functions correspond to the expected payoff function of the

[22]Sequential move games in which the second mover observes the same signal independent of the first mover's action are strategically equivalent to simultaneous-move games; see Hart (1992) for details.

algorithmic player and the short-run player, respectively, conditional on the state $s$. Let $a^*(s) = (a_Q^*(s), a_{\text{SR}}^*(s))$ be a strict Nash equilibrium of the game $G(s)$.

**Theorem 2** (Transparent algorithm)**.** *Suppose the signalling structure $(\Phi, p)$ has full support. Suppose the algorithm is transparent.*
*Then any strict Nash equilibrium $a^*(s)$ is an asymptotically stable outcome. Formally, for any $\xi > 0$ there exists $K \in \mathbb{N}$ and, for each $a_Q \in A_Q$, an open neighborhood $\mathcal{O}_{a_Q}$ around $\mathbb{E}[u_Q(a_Q, a_{\text{SR}}^*(s), \omega)|s]$ such that,*

$$\mathbb{P}\left[\lim_{T\to\infty} \frac{\sum_{t=K}^{K+T} \mathbb{1}\{(a_Q^t, a_{SR}^t, s^t) = (a_Q^*(s), a_{SR}^*(s), s)\}}{\sum_{t=K}^{K+T} \mathbb{1}\{s^t = s\}} = 1 \;\middle|\; \forall a_Q,\, Q_K(s, a_Q) \in \mathcal{O}_{a_Q}\right] \geq 1 - \xi.$$

*Proof in Appendix A.2.*

**Corollary 1.** *Under the hypotheses of Theorem 2, for any sequence of updating rates $(\alpha^t)$ and experimentation rates $(\varepsilon^t)$, there exists $K$ and an open set of initial $Q$-values $Q^0(s, a_Q), a_Q \in A_Q$, such that*

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

*when the algorithm's parameters are $\langle Q^0, (\alpha^{K+t}), (\varepsilon^{K+t})\rangle$.*

Theorem 2 should be interpreted as stating that Nash equilibria are asymptotically stable.[23] More precisely, Theorem 2 states that the vector $(\mathbb{E}[u_Q(a_Q, a_{\text{SR}}^*(s), \omega)|s])_{a_Q \in A_Q}$ — the expected payoff when the short-run player plays according to the Nash equilibrium action $a_{\text{SR}}^*(s)$ — is a stochastically absorbing state in the space of $Q$-values. If the $Q$-values in some late enough period $K$ are close to the absorbing state, the probability that the $Q$-values stay in the same neighborhood is close to 1. If the $Q$-values remain in this neighborhood, the algorithm's greedy action is to play according to the Nash equilibrium. Theorem 2 states that, once this neighborhood is reached, with probability at least $1 - \xi$, the fraction of periods in which state $s$ occurs and the Nash equilibrium $a^*(s)$ is played (by both the algorithmic player and the short-run player) converges to 1.

Theorem 2 is an asymptotic result in two ways. First, experimentation rates need to be sufficiently low so that they do not affect the short-run players' best-responses. Since experimentation rates vanish, this is always satisfied after some period. Second, the updating parameters need to be small enough. If updating rates are close to 1, a single payoff observation moves the $Q$-values by a lot. Thus,

[23] Theorem 5 in the Online Appendix provides a generalization of the result for $S_Q \neq S_{\text{SR}}$.

updating needs to be slow enough to guarantee that the $Q$-values remain contained in a neighborhood with high probability.

It is not sufficient to assume that the algorithm's greedy action corresponds to the Nash equilibrium action $a_Q^*(s)$ for the conclusion of Theorem 2 to obtain. More precisely, it does not suffice that the $Q$–value assigned to the equilibrium action $a_Q^*(s)$ is the highest $Q$–value at some period $K$. To see this, suppose the $Q$–values for all actions are higher than the Nash payoff $\mathbb{E}[u_Q(a_Q^*(s), a_{SR}^*(s), \omega)|s]$. Then the $Q$–values for all actions decrease (in expectation) while $a_Q^*(s)$ is played. Consequently, the $Q$-value for $a_Q^*(s)$ may drop below the $Q$–value of some other action: play moves away from the Nash equilibrium.

Theorem 2 is silent about the long-run payoffs attained by the algorithm. However, conditional on the event that the $Q$-values lie in the neighborhood $\mathcal{O}_{a_Q}$ for all $a_Q$ in some period $t \geq K$, the following holds: the long-run average (expected) payoff of the algorithm in state $s$ belongs to the interval

$$[(1-\xi)\mathbb{E}[u_Q(a_Q^*(s), a_{\text{SR}}^*(s), \omega)|s] - \xi M, (1-\xi)\mathbb{E}[u_Q(a_Q^*(s), a_{\text{SR}}^*(s), \omega)|s] + \xi M],$$

where $M$ is an upper bound on the norm of the algorithmic player's expected payoff in state $s$, $\mathbb{E}[u_Q(\cdot, \cdot, \omega)|s]$.

What is the intuition behind Theorem 2? When the algorithm is transparent, the short-run player in period $t$ knows the history of the states, the algorithm's actions and its realized payoffs before period $t$, i.e., $\text{SR}^t$ observes $h^{t-1} \in (S_Q \times A_Q \times U_Q)^{t-1}$. The $Q$-values in period $t$, $Q^t(s, a_Q), s \in S_Q, a_Q \in A_Q$, are determined uniquely by the history $h^{t-1}$. Hence, $\text{SR}^t$ knows the $Q$-values in period $t$. Since $\text{SR}^t$ observes the algorithm's state in period $t$ as well, $s_Q^t \in S_Q$, $\text{SR}^t$ knows the action the algorithm takes, unless it experiments.

When the signals are noisy and the experimentation probability is low enough, the short-run player believes, with probability close to 1, that the algorithmic player has chosen the greedy action, irrespective of the signal observed.[24] Consequently, the short-run player plays a best-response to the algorithm's greedy action, independently of the signal received and thus independently of the action the algorithm has actually played. In other words, the short-run players play a best-response against the expected action of the algorithm and ignore the signal $\phi$.

When the greedy action corresponds to a Nash equilibrium action of the auxiliary game $G(s)$, the short-run players' best-response is to play according to the Nash equilibrium as well. Hence, as long as the algorithm's greedy action does not change and the short-run players ignore the signal $\phi$, play is according to this

[24]The observation that imperfect signals convey little information when an action is chosen with probability (close to) 1 is ubiquitous in economics and game theory. Prominent examples are Bagwell's (1995) study of the first-mover advantage or repeated games with private monitoring (e.g., Matsushima (1991) and Bhaskar and Damme (2002)).

Nash equilibrium (up to experimentation). Moreover, when the the short-run players keep playing $a^*_{\text{SR}}(s)$, the $Q$-values converge to the corresponding expected payoff for the algorithm, $Q^t(s, a_Q) \to \mathbb{E}[u_Q(a_Q, a^*_{\text{SR}}(s), \omega)|s]$, for each $a_Q \in A_Q$. If the $Q$-values are in a neighborhood of this limit already, they remain in the neighborhood with high probability, provided that updating rates are sufficiently small.[25]

It is instructive to contrast the intuition behind Theorem 2 with the reasoning behind Theorem 1. The positive result under perfect monitoring obtains because the short-run players play a best-response to the *actual* action of the algorithmic player. Failure to converge to the Stackelberg outcome is due to the short-run players playing a best-response to the *expected* action. Consequently, the payoff the algorithm receives in period $t$ depends not only on its actual, sampled action but on the history. Therefore, convergence to the Stackelberg outcome fails.

The assumptions on the signalling structure in Theorem 2 are mild. Full support signalling includes the important case of simultaneous-move games. However, it also holds for signalling structures that are close to perfect signalling.[26] Theorem 2 can thus be viewed as a negative result: no matter how precise the signalling structure is, Nash equilibria of the auxiliary game are asymptotically stable. Consequently, there is no guarantee that the Stackelberg outcome will obtain for all parameters of the algorithm. The positive result for perfect signalling, Theorem 1, is, therefore, a knife-edge case.

## 3.3. Opaque algorithm

In this section, we discuss the case of opaque algorithms. Recall that the algorithm is opaque if the short-run players do not observe its inputs. We show that the results differ starkly from the ones obtained for transparent algorithms: when the algorithm's information about the realized shock and the short-run players' signal about the algorithm's action are sufficiently precise, the $Q$-learning algorithm learns to play the Stackelberg action.

**Example 1** (continued)**.** Suppose that the algorithm is opaque. The consumer neither observes past quality choices nor the current state of the economy. Suppose the consumer observes the service quality almost perfectly: with probability close to 1, the consumer sees a signal $h$ when the retailer provides high quality and a different signal $l$ when the retailer provides low quality.

[25] If the algorithmic player's payoff is deterministic conditional on the state $s$, then the $Q$-values remain in the neighborhood with probability 1.

[26] Theorem 2 remains true if the order of moves is reversed: suppose that in each period $t$, the short-run player $\text{SR}^t$ chooses an action, the algorithm observes a signal about the action and then chooses its action. In that case, the signals can even be perfect. Hence, Nash equilibria (of the auxiliary simultaneous move game) are stable under a wide range of conditions.

As in the case with a transparent algorithm, in states $s_1$ and $s_3$ the algorithm eventually learns to play the best action, i.e., to provide high–quality service in state $s_1$ and to provide low–quality service in state $s_3$. What happens in state $s_2$? The algorithm's and consumer's behavior in the beginning depend on the algorithm's initial guess and the experimentation probabilities. Once the consumer believes that learning in states $s_1$ and $s_2$ has occurred with probability close to 1, the consumer expects the retailer to provide high-quality service in state $s_1$ and low-quality service in state $s_3$. Consequently, the consumer believes the retailer provides high–quality service with probability at least $q(s_1)$ and low-quality service with probability at least $q(s_3)$, irrespective of the experimentation probabilities.

If the signals the consumer receives are precise enough, the consumer believes that the service quality is high with probability close to 1 after observing signal $h$; after observing signal $l$, she believes the service quality is high with probability close to 0. Crucially, this is true irrespective of the retailer's behavior in state $s_2$. Because the retailer keeps providing high quality in state $s_1$ and low quality in state $s_3$, up to experimentation, the signals convey enough information in all future periods. The consumer then behaves as follows: return the product if signal $h$ is observed and ask for a repair if the signal $l$ is observed.

Given this behavior of the consumer, the algorithm eventually attaches a higher average payoff to providing high–quality service than to providing low quality in state $s_2$. Because the signal is precise, the retailer obtains a payoff close to 2 — the payoff the retailer would have gotten if it had complete information about the payoffs and if it could commit to an action. Hence, if the algorithm is opaque, it learns to play according to full information, full commitment benchmark in every state. ■

When the algorithm is opaque, it learns the Stackelberg action under two conditions. First, it has (payoff-relevant) information about the realized shock $\omega$ that the short-run players do not have. Second, the short-run players observe a precise enough signal about the algorithm's action. The next two definitions make these notions formal.

**Definition 6** (Rich Partitions)**.** *A finite partition $S$ of $\Omega$ is a* rich partition *if for every $\omega \in \Omega$ there exists $s \in S$ such that*

$$u_Q(a, \omega) \geq u_Q(a', \omega) \iff \mathbb{E}[u_Q(a, \tilde{\omega})|\tilde{\omega} \in s] \geq \mathbb{E}[u_Q(a', \tilde{\omega})|\tilde{\omega} \in s]$$

*for every two action profiles $a, a' \in A_Q \times A_{\text{SR}}$.*

A *rich partition* $S$ requires the following: for every ordinal preference relation over actions $A_Q \times A_{\text{SR}}$ induced by the payoff function $u_Q(\cdot, \omega)$ for some realization $\omega$, there exists a cell $s$ in the information partition $S$ such that the expected payoff

function conditional on $s$ represent the same ordinal preferences. A rich partition exists for all $\Omega$ because action sets are finite.

We next introduce a formal notion of the precision of the signals $\phi$ about the algorithmic player's action.

**Definition 7** ($\gamma$-perfect signalling)**.** *The signalling structure* $(\Phi, p)$ *is* $\gamma$-perfect *if* [27]

1. *for each* $a_Q$, $p(\cdot|a_Q)$ *has full support;*
2. *for each* $a_Q$, *there exists* $\phi_{a_Q}$ *such that* $p(\phi_{a_Q}|a_Q) \geq 1-\gamma$.

The main result for opaque algorithms is the following theorem.

**Theorem 3** (Opaque algorithm)**.** *Suppose the signalling structure* $(\Phi, p)$ *has full support. Suppose the algorithm is opaque and the algorithm's state space* $S_Q$ *is a rich partition.*
*When the signals about the algorithm's action are precise enough, 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* $\gamma$*-perfect monitoring structures* $(\Phi, p)$ *with* $\gamma \leq \overline{\gamma}$, *for every state* $s \in S_Q$,

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

*and*

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

*almost surely for all parameters* $\langle Q^0, (\alpha^t), (\varepsilon^t)\rangle$ *of the algorithm. Here,* $M$ *is a bound on the norm of the algorithmic player's expected payoff.*
*Proof in Appendix A.3.*

**Remark 1.** *The hypothesis that* $S_Q$ *is a rich partition can be weakened. It suffices to assume instead that for every action* $a_Q$ *there exists a state* $s_{a_Q} \in S_Q$ *such that*[28]

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

[27] The second condition can be replaced by: for each $a_Q$ there exists $\Phi_{a_Q} \subset \Phi$ such that $p(\Phi_{a_Q}|a_Q) \geq 1-\gamma$ and $p(\Phi_{a_Q}|a'_Q) \leq \gamma$ for all $a'_Q \neq a_Q$.

[28] The Online Appendix discusses sufficient and necessary conditions under which the algorithm converges to playing the Stackelberg action. The condition in equation (1) cannot be weakened to strict dominance; see Example 6 in the Online Appendix.

■

Theorem 3 makes two statements. First, it states that an opaque algorithm learns to play the Stackelberg action in each state $s$: the fraction of periods in which the algorithm plays according to the Stackelberg action converges to 1 almost surely. Second, Theorem 3 states that an opaque algorithm achieves a long-run payoff close to the Stackelberg payoff in every state. In the long-run, the payoff the algorithmic player obtains is approximately equal to the highest payoff it could obtain under full information about the payoff functions. When the algorithm is opaque, the two main properties of $Q$-learning in single-player environments — learning the optimal action and achieving the maximal payoff — carry over to the strategic environment with short-lived, best-responding players!

Theorem 3 relies on two hypotheses. First, the algorithm's information about the realized $\omega$ is precise enough: the information partition $S_Q$ is rich according to Definition 6. This implies, in combination with Assumption 3, that the algorithm has enough payoff-relevant information that the short-run players do not have. This is needed to guarantee that the short-run player's expect the algorithm to play each action with a probability that is independent of its parameters and strictly positive.

Second, the signals the short-run players receive about the algorithm's action must be precise enough. Learning to play the Stackelberg action requires that the short-run players' behavior depends on the action sampled by the algorithm. Precise signals are needed to ensure that the short-run players' action is sensitive to the action taken by the algorithm. However, as we argue in Section 3.2, precise signals alone are not sufficient to ensure learning. It is the combination of a rich information partition of the algorithm and precise signals that ensure learning of the Stackelberg action when the algorithm is opaque.

The proof of Theorem 3 consists of three main steps. In the first step, we argue that for every action $a_Q$ there is a state $s_{a_Q} \in S_Q$ such that the worst possible expected payoff for the algorithm is strictly higher when playing $a_Q$ than the highest possible expected payoff when playing any other action $a'_Q$. We then show that for any strategy of the short-run players, the $Q$-value in that state $s_{a_Q}$ is eventually highest for the action $a_Q$. Since for any action $a_Q$ there exists such a state, there is a period $T$ such that the short-run players expect the algorithm to play each action with a minimal probability $\underline{q}$ in every period $t \geq T$. In particular, this minimal probability does not depend on the experimentation rates $(\varepsilon^t)$ or other parameters of the algorithm.

In the second step we show that, eventually, the short–run players' strategies become close to stationary. If the signalling structure is precise enough, there is a set of signals $\Phi'$ such that the action the short-run players take after observing a signal in $\Phi'$ does not depend on the period $t$, provided $t \geq T$. This makes the

short-run players' strategies close to stationary.

In the third step, we use the approximate stationarity to provide bounds on the long-run trajectories of the $Q$–values. Watkin's Theorem does not apply because the short-run players' strategy is not stationary.[29] However, we can show that, with probability 1, for each state-action combination $(s, a_Q)$, the corresponding $Q$-value is eventually contained in an interval. If signals are precise, those intervals are small enough such that, for each state, the action $a_Q^{\text{Stack}}(s)$ that attains the Stackelberg value has the highest $Q$-value. Consequently, with probability 1, the greedy action corresponds to the action attaining the Stackelberg value. Play of the algorithm converges, up to experimentation.

We remark that the strategies of the short-run players converge as well. There exists a period $\tilde{T}$ such that $\sigma_{\text{SR}^t} = \sigma_{\text{SR}^{\tilde{T}}}$ for all periods $t \geq \tilde{T}$. However, the short-run players' strategies converge only after they believe, with probability close to 1, that the algorithm's greedy action has converged in every state $s$.

Precise signals about the algorithm's actions play two distinct roles. First, they ensure that the short-run players' strategy eventually becomes close to stationary. This then ensures that the $Q$-values converge to an interval, even if they do not converge to a single point. Second, precise signals ensure that the short-run players eventually best-respond to the sampled action with high probability. This ensures that play of the algorithm converges. Moreover, best-responding with probability close to 1 ensures that the algorithm learns to play according to the Stackelberg outcome.

The precision of the signalling structure, $\overline{\gamma}$, in the statement of Theorem 3 does not depend on the parameters of the algorithm. It depends on the (expected) payoff functions of the algorithmic player and the short-run players, the probability distribution of the random variable $\omega$, and the algorithm's state space $S_Q$.

## 3.4. Comparison between transparent and opaque algorithms

In this section, we compare the long-run outcomes for transparent and opaque algorithms. We first focus on whether the algorithm learns to play according to the Stackelberg outcome and on the long-run payoffs the algorithm obtains. We then compare the long-run payoffs of the short-run players when facing a transparent algorithm and when facing an opaque algorithm.

Recall that the short-run players' information about the shock $\omega$, given by the information partition $S_{\text{SR}}$, differs when the algorithm is transparent and when the algorithm is opaque: $S_{\text{SR}} = S_Q$ in the former case and $S_{\text{SR}} = \{\Omega\}$ in the latter case. Consequently, to make the comparison meaningful while ensuring that Assumption

[29] The $Q$-values need not converge if the action the short-run players take depend on the period $t$.

2 holds, we assume here $\mathbb{E}[u_{\text{SR}}(\cdot,\omega)|s] = \mathbb{E}[u_{\text{SR}}(\cdot,\omega)|s']$ for all $s, s' \in S_Q$.

### 3.4.1. Algorithm's learning and payoffs

Sections 3.2 and 3.3 show how the outcomes for transparent and opaque algorithms differ both with regards to learning and with regards to payoffs. When the algorithm is transparent, it need not learn to play according to the Stackelberg outcome. Consequently, the expected long-run payoff may be lower than the Stackelberg payoff. In contrast, an opaque algorithm learns to play according the Stackelberg action and obtains a long-run payoff close to the Stackelberg payoff. The algorithm performs better when opaque than when transparent.

The algorithm thus achieves higher payoffs when the short-run players have less information about its inputs. The reason is that the information short-run players have about the algorithm affects their behavior and consequently the algorithm's learning. When the short-run players have less information about the algorithm's inputs, learning the Stackelberg outcome obtains because the short-run players condition their action on the signal they receive about the algorithm's action. Hence, the short-run players' behavior remains responsive to the signals they see about the algorithm's action. In contrast, when the short-run players know the algorithm's inputs, their behavior eventually becomes independent of these signals and thus independent of the action taken by the algorithm.

We remark that the algorithm does not benefit from the short-run players' ignorance about the state because it exploits this ignorance. By Assumption 2, all information about the short-run players' payoff that is contained in state $s \in S_Q$ is available to the short-run players as well. Hence, even if informed about the algorithm's state before making her decision — keeping her beliefs about the algorithm's action unchanged — the short-run player would not alter her choice.

Moreover, the algorithm's profit when opaque are not higher because the short-run players fail to best-respond to the action the algorithm has actually chosen. Under the conditions of Theorem 3, the short-run players eventually best-respond to the action played by the algorithm with probability at least $1-\gamma$. Hence, any payoff gains or losses that are due to the short-run players failing to best-respond vanish as signals become precise.

The role asymmetric information about the realized state plays is different. If the algorithm observes the state but the short-run player does not, the signals about its action remain informative. The short-run players' uncertainty about the state prevents them from ignoring the signals.[30] Hence, the signals convey enough

[30] Maggi (1999) finds a similar effect in a sequential quantity setting game. When the first-mover's cost of production is her private information and signals are sufficiently precise, equilibrium payoffs are close to the Stackelberg payoff. There are important differences to our results. First, Maggi's (1995) result is an equilibrium analysis. In our setting, play need not converge to an

information to affect the short-run players' behavior. Eventually, the action the short-run player takes depends on the signal received. This enables the algorithm to learn the Stackelberg action in every state.

Opaque algorithms perform better than transparent algorithms not only with regards to learning, but also with regards to the expected long-run payoffs. As the signals $\phi$ become precise, the long-run expected payoffs for the opaque algorithm are higher than for the transparent algorithm, irrespective of the algorithm's parameters. Proposition 1 makes this relation precise.

For fixed parameters of the algorithm, let

$$W^{Q,x}_{(\Phi,p)} = \liminf_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[u_Q\left(a^t_Q, a^t_{\mathrm{SR}}, \omega^t\right)\right]$$

be the long-run average expected payoff of the algorithm. Here, $x = \mathrm{tra}, \mathrm{opa}$ denotes whether the algorithm is transparent or opaque.

**Proposition 1.** *Suppose the signalling structure $(\Phi, p)$ has full support. Suppose for each $a_Q \in A_Q$ there exists $s \in S_Q$ such that (1) is satisfied. Let $(\Phi, p_n)$ be a sequence of $\gamma_n$-perfect signalling structures with $0 < \gamma_n \to 0$. Then the following hold:*

1. *For any parameters $\langle Q^0, (\alpha^t), (\varepsilon^t)\rangle$ of the algorithm,*

$$\limsup_{n\to\infty} W^{Q,tra}_{(\Phi,p_n)} \leq \lim_{n\to\infty} W^{Q,opa}_{(\Phi,p_n)} = \mathbb{E}\left[u^{\mathrm{Stack}}_Q(s)\right],$$

   *where the expectation is over the state $s \in S_Q$.*

2. *Suppose the algorithm's state space $S_Q$ is a rich partition. Unless the short-run player has a strictly dominant action, for any parameters of the algorithm,*

$$W^{Q,tra}_{(\Phi,p_n)} < \mathbb{E}\left[u^{\mathrm{Stack}}_Q(s)\right]$$

   *for all $n$.*

3. *Suppose the algorithm's state space $S_Q$ is a rich partition. Unless the short-run player has a strictly dominant action, for any $n$ large enough there exists*

equilibrium of the auxiliary stage game. Moreover, Maggi (1999) assumes a continuum of actions, states and signals as well as a specific payoff function. In our finite setting, private information about the state and precise signals are not sufficient to guarantee convergence to the Stackelberg outcome; see Example 6 in the Online Appendix for a striking example.

*parameters of the algorithm such that*

$$W^{Q,tra}_{(\Phi,p_n)} \leq W^{Q,opa}_{(\Phi,p_n)} - \delta$$

*for a constant* $\delta > 0$ *independent of* $(\Phi, p_n)$ *and the algorithm's parameters.*

*Proof in Appendix A.4.*

The hypothesis that for each action $a_Q$ there exists a state $s$ such that (1) holds ensures that the algorithm learns the Stackelberg action when the signals are precise enough; see Remark 1.

The first part of Proposition 1 states that the algorithmic player's long-run expected payoff is higher when the algorithm is opaque than when the algorithm is transparent as signals become arbitrarily precise. Theorem 3 shows that the opaque algorithm achieves the Stackelberg payoff in each state as signals become perfect. However, the transparent algorithm need not learn the Stackelberg action and thus achieves a lower payoff.

We focus on the limit payoffs as signalling becomes perfect to disentangle two effects on the algorithm's payoff: first, whether the algorithm learns to play the Stackelberg action; second, whether short-run players best-reply to the algorithm's action. When the algorithm is opaque and signals are imperfect, the short-run players play an action different from their best-response with positive, albeit small, probability. Focusing on the limit as signals become perfect eliminates this second effect – which can be positive or negative.

The second part of Proposition 1 states that for any fixed signalling structure and any parameters of the algorithm, a transparent algorithm achieves a payoff strictly less than the (expected) Stackelberg payoff; the algorithm fails to learn the Stackelberg action with positive probability. The result holds under two conditions. First, the short-run players do not have a strictly dominant action. If the short-run players have a strictly dominant action, they play this action in every period and after every history. Consequently, the algorithm faces a stationary environment, and asymptotically achieves the Stackelberg payoff. The second condition is that the algorithm's state space is a rich partition. This condition rules out that a transparent algorithm asymptotically plays the Stackelberg action with probability 1 in each state.

The first and the second part of Proposition 1 hold for arbitrary parameters of the algorithm. The third part considers long-run payoffs when the algorithm's parameters are chosen adversely. It states that, for some parameters of the algorithm, a transparent algorithm attains a payoff strictly below the payoff an opaque algorithm attains. Recall that, when the algorithm is opaque, its long-run payoff does not depend on its parameters. When the algorithm is transparent, there exists parameters such that it fails to learn the Stackelberg action and thus

achieves a payoff strictly below the Stackelberg payoff; see Theorem 2 and its Corollary. Consequently, the transparent algorithm achieves a payoff strictly below the expected Stackelberg payoff, and thus less than the long-run payoff an opaque algorithm would achieve.

### 3.4.2. Short-run players' payoff

**Example 1** (continued)**.** Consider again Example 1. Recall that in states $s_1$ and $s_3$, the algorithm learns to play the Stackelberg action when it is transparent and when it is opaque. The opaque algorithm learns the Stackelberg action in state $s_2$ as well. In contrast, a transparent algorithm need not learn the Stackelberg action in state $s_2$. The payoff for the consumer is maximized when she asks for a return and the algorithm provides high quality. Hence, consumers benefit when the algorithm learns to play the Stackelberg action in state $s_2$. As a consequence, consumers receive a higher average payoff when signals are precise and the algorithm is opaque than when the algorithm is transparent. ■

How does the short-run players' payoff depend on whether the algorithm is transparent or opaque? Example 1 shows that the short-run players can, on average, receive a higher payoff when the algorithm is opaque than when the algorithm is transparent. However, this does not hold for all games. Under what conditions are the short-run players, on average, better off when either all short-run players know the algorithm's inputs or when no short-run players has information about the algorithm's inputs? Put differently, under what conditions are the short-run players better off with an opaque algorithm compared to a transparent algorithm?

As a measure of the short-run players' average payoff, we take the long-run expected average of all short-run players' payoffs, i.e., we consider

$$W^{\mathrm{SR},x}_{(\Phi,p)} = \liminf_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[u_{\mathrm{SR}}\left(a^t_Q, a^t_{\mathrm{SR}}, \omega^t\right)\right].$$

Observe that each single short-run player $\mathrm{SR}^t$ prefers having more information about the algorithm's inputs. This holds for two reasons. First, there is a single short-run player in each period. Second, the action taken by $\mathrm{SR}^t$ affects only the short-run players of later periods, but leaves the strategy of short-run players that precede $\mathrm{SR}^t$ unchanged. Therefore, changing the information $\mathrm{SR}^t$ has does not affect the action chosen by the algorithm in period $t$. Thus, $\mathrm{SR}^t$ faces, essentially, a single-player decision problem, and having more information increases the payoff of $\mathrm{SR}^t$.

However, changing the information of $\mathrm{SR}^t$ affects the action she takes. This affects

the payoff the algorithm receives in period $t$ and thus its $Q$-values in subsequent periods. There is thus an externality of $\text{SR}^t$'s information on later short-run players through the algorithm's learning.

Having no information about the algorithm's inputs can increase the short-run players' average payoff if the externality is positive. Sufficient for the short-run players' average payoff to be higher when the algorithm is opaque than when the algorithm is transparent is that the short-run players benefit from the algorithm playing the Stackelberg action in each state. The first part of Proposition 2 makes this precise.

**Proposition 2.** *Suppose the signalling structure $(\Phi, p)$ has full support. Suppose for each $a_Q \in A_Q$ there exists $s \in S_Q$ such that (1) is satisfied. Let $(\Phi, p_n)$ be a sequence of $\gamma_n$-perfect signalling structures with $0 < \gamma_n \to 0$. Then the following hold:*

1. *Suppose that for each $s \in S_Q$ such that (1) is not satisfied for any $a_Q$, the Stackelberg action $a_Q^{\text{Stack}}(s)$ maximizes the short-run player's payoff, i.e.,*

$$\mathbb{E}\left[u_{\text{SR}}\left(a_Q^{\text{Stack}}(s), \text{BR}\left(a_Q^{\text{Stack}}(s)\right), \omega\right)\right] \geq \mathbb{E}\left[u_{\text{SR}}\left(a_Q, \text{BR}(a_Q), \omega\right)\right] \forall a_Q.$$

   *For any parameters of the algorithm $\langle Q^0, (\alpha^t), (\varepsilon^t)\rangle$,*

$$\limsup_{n\to\infty} W^{\text{SR},tra}_{(\Phi,p_n)} - W^{\text{SR},opa}_{(\Phi,p_n)} \leq 0.$$

2. *Suppose that for each $s \in S_Q$ for which (1) is not satisfied for any $a_Q$, the auxiliary game $G(s)$ has a pure Nash equilibrium that is Pareto-dominated by the Stackelberg outcome $(a_Q^{\text{Stack}}(s), \text{BR}(a_Q^{\text{Stack}}(s)))$. Suppose there is at least one such state $s$. For any $n$ large enough there exist parameters of the algorithm such that*

$$W^{\text{SR},tra}_{(\Phi,p_n)} \leq W^{\text{SR},opa}_{(\Phi,p_n)} - \delta$$

   *for a constant $\delta > 0$ independent of $(\Phi, p_n)$ and the algorithm's parameters.*

*Proof in Appendix A.5.*

Proposition 2 imposes conditions on both the algorithmic player's payoffs and on the short-run players' payoff function. The condition on the algorithmic player's payoffs guarantee that an opaque algorithm learns the Stackelberg action, provided signals are precise enough. The conditions on the short-run player's payoffs ensure that the short-run players benefit from the algorithm playing the Stackelberg action.

In a state $s$ such that equation (1) holds for some action $a_Q$, the algorithm plays the Stackelberg action asymptotically, irrespective of the short-run players'

behavior. Consequently, the payoff for the short-run players in such a state does not depend on whether the algorithm is transparent or opaque.

The first part of Proposition 2 says that the short-run players receive a higher average payoff when the algorithm is opaque as signals become precise when the Stackelberg outcome maximizes the short-run player's payoff. When the Stackelberg outcome maximizes the short-run player's payoff, the short-run players' benefit from the algorithm's learning of the Stackelberg action. Consequently, the short-run players receive a higher payoff when the algorithm is opaque so that it learns to play the Stackelberg action. This result holds irrespective of the parameters of the algorithm.

The second part of Proposition 2 provides sufficient conditions such that, for some parameters of the algorithm, the short-run players' average payoff is higher when the algorithm is opaque. By Theorem 2 and its Corollary, play can get stuck at a Nash equilibrium of the auxiliary game when the algorithm is transparent. If this Nash equilibrium is Pareto-dominated by the Stackelberg outcome, the short-run players receive a higher payoff when the algorithm learns the Stackelberg action. Hence, they receive a higher payoff when the algorithm is opaque. As a result, the conditions in the second part of Proposition 2 ensure that short-run players receive a higher average payoff with an opaque algorithm compared to a transparent one when the algorithm's parameters are selected to minimize their payoff.[31]

**Example 2.** *Product recommendation.* Suppose the algorithmic player is an intermediary who recommends products to a prospective consumer. Products have two possible characteristics, $a, b$, that determine the consumer's valuation; suppose that the consumer's valuation for the characteristics satisfies $v_a < 0 < v_b$. The intermediary can distinguish between the characteristics, and recommends a product to the consumer. The consumer sees the product, but cannot perfectly infer its characteristics.

If the consumer buys the product, her payoff equals its valuation. When making a sale, the intermediary receives a fixed commission. If the consumer does not make a purchase, she and the intermediary receive a payoff of 0. Moreover, assume that there are states such that a product with characteristic $a$ or $b$ is not available.[32]

The actions of the intermediary are "recommend product with characteristic $a$" and "recommend product with characteristic $b$". The consumer's best response to

[31]The conditions in the second part of Proposition 2 are not satisfied if the algorithm's state space is a rich partition. In this case, the proposition remains true if the probability of states such that the auxiliary game $G(s)$ has a pure Nash equilibrium that is Pareto-dominated by the Stackelberg outcome is close enough to 1.

[32]Suppose that the intermediary's profit from recommending a product that is not available is negative.

the intermediary's action is to buy if the intermediary recommends the product with characteristic $b$, and not to buy otherwise. The Stackelberg outcome is as follows. The intermediary recommends the product with characteristic $b$ if this product is available, and recommends the product with characteristic $a$ if this is the only available product. The consumer buys the product if the intermediary recommends the product with characteristic $b$. ■

## 3.5. Comparison to a strategic long-lived player

In this section, we ask whether the long-run Player 1 suffers or benefits from using a $Q$-learning algorithm. We compare the payoffs the long-run player can achieve when playing according to the algorithm to the payoffs he can achieve when behaving optimally.[33] Assume that the long-run Player 1 is not restricted to using an algorithm. Instead, Player 1 can use history-dependent strategies. We compare the equilibrium payoffs such a strategic long-lived player can obtain to the payoffs a $Q$-learning algorithm can attain.

Suppose the long-lived player LR is not restricted to using a $Q$-learning algorithm. Instead, he behaves strategically, in a way to maximize his own payoff in a repeated game with the short-run players. Assume both the long-run player LR and the short-run players ($\mathrm{SR}^t$) know the payoff functions $(u_Q(\cdot,\omega) \equiv u_{\mathrm{LR}}(\cdot,\omega), u_{\mathrm{SR}}(\cdot,\omega))$ and the probability distribution $q(\cdot)$ over $\Omega$.[34]

A strategy for the long-run player LR is a map

$$\sigma_{\mathrm{LR}} : \bigcup_{t=0,1,\ldots} (S_Q \times A_Q \times U_Q)^{t-1} \times S_Q \to \Delta(A_Q).$$

[33] It is not our goal to explain why firms might use such learning algorithms, as restrictive as they might be. Aside from conceptual reasons (e.g., a prior-free instead of a Bayesian learning procedure), there are computational reasons (complexity and speed) that make them attractive in practice.

[34] The results remain unchanged if we instead assume that the long-run player does not know the payoff functions. To be more precise, consider the following setting. Suppose for each $\omega$ there are finitely many possible payoff functions $\{u_Q^k,(\cdot,\omega), u_{\mathrm{SR}}^k(\cdot,\omega)\}_{k\in K}$. Fix a prior probability over the set of payoff functions $\rho(k)$. Assume the short-run players observe the realized payoff functions, i.e., the realized $k$, but the long-lived player does not. This defines a repeated game of incomplete information.
Then the following holds. Let $u_{\mathrm{LR}}^k$ be a sequential equilibrium payoff for the long-run player in the repeated game when payoff functions are known. Then there exists a sequential equilibrium of the repeated game of incomplete information such that the long-run player's payoff is $u_{\mathrm{LR}}^k$ when $\{u_Q^k,(\cdot,\omega), u_{\mathrm{SR}}^k(\cdot,\omega)\}$ are the realized payoff functions.
This equivalence holds because we consider a patient long-term player who assesses payoffs based on a limit-of-means criterion. By focusing on limit-of-means payoffs rather than discounted payoffs, the emphasis shifts from what is optimal to learn in equilibrium to what can actually be learned in equilibrium.

For the strategies of the short-run players, we distinguish two main cases: whether the short-run players observe the history of play or not. In the former case, the short-run players observe the past states, past actions of the long-run player and his past realized payoffs, in addition to their information about the realization of the shock $\omega$ and a signal $\phi$ about the long-run player's action. Formally, a strategy of player $\text{SR}^t$ is a map[35]

$$\sigma_{\text{SR}^t} : (S_Q \times A_Q \times U_Q)^{t-1} \times S_{\text{SR}} \times \Phi \to \Delta(A_{\text{SR}}).$$

In the second case, the short-run players only observe their information about the realization of the shock $\omega$ and a signal $\phi$ about the long-run player's action. Formally, a strategy of player $\text{SR}^t$ is a map[36]

$$\sigma_{\text{SR}^t} : S_{\text{SR}} \times \Phi \to \Delta(A_{\text{SR}}).$$

Given a strategy profile for each player $(\sigma_{\text{LR}}, (\sigma_{\text{SR}^t}))$, the payoff to the long-run player is

$$U(\sigma_{\text{LR}}, (\sigma_{\text{SR}^t})) = \liminf_{T\to\infty} \sum_{t=0}^{T-1} \mathbb{E}^{(\sigma_{\text{LR}}, (\sigma_{\text{SR}^t}))} \left[ u_Q(a^t_{\text{LR}}, a^t_{\text{SR}}, \omega^t) \right].$$

This defines a repeated game. We focus on sequential equilibria in pure strategies of this game which we refer to as equilibria for short.[37]

The next theorem compares the equilibrium payoffs of the strategic long-lived player LR with the payoffs a $Q$-learning algorithm can obtain.

**Theorem 4.** *Suppose the signalling structure $(\Phi, p)$ has full support.*
*Let $v^*_{\text{LR}}$ be the highest equilibrium payoff the strategic long-run player can achieve.*
*Let $v^*_Q$ be the highest payoff the $Q$-learning algorithm can achieve for any parameters $\langle Q^0, (\alpha^t), (\varepsilon^t) \rangle$.*[38]

[35] When $S_Q = S_{\text{SR}}$, this case corresponds to the case of a transparent algorithm according to Definition 1.

[36] When $|S_{\text{SR}}| = 1$, this case corresponds to the case of an opaque algorithm according to Definition 2.

[37] When playing against an algorithm, the short-run players do not play a fully mixed strategy in any equilibrium for generic payoff functions and parameters. Allowing for mixed strategy equilibria in the game with a strategic long-run player would distort the comparison. We provide sufficient conditions for the existence of a pure strategy equilibrium in Appendix A.6.

[38] Consider the following game. Suppose the algorithmic player knows the payoff functions and the probability distribution $q$. The algorithmic player chooses parameters $\langle Q^0, (\alpha^t), (\varepsilon^t) \rangle$ under the restriction that $(\alpha^t)$ satisfies Assumption (Step-Size) and $(\varepsilon^t)$ satisfies Assumption (Experimentation). The choice of parameters is observed by the short-run players. Then the short-run players play according to a Nash equilibrium of the resulting subgame. Then the following

1. *Suppose the short-run players observe the outcome of past interactions, i.e., the strategy for* $\mathrm{SR}^t$ *is a map*

$$\sigma_{\mathrm{SR}^t} : (S_Q \times A_Q \times U_Q)^{t-1} \times S_{\mathrm{SR}} \times \Phi \to A_{\mathrm{SR}}.$$

*Then* $v^*_{\mathrm{LR}} \geq v^*_Q$.

2. *Suppose the short-run players do not observe the outcome of past interactions, i.e., the strategy for* $\mathrm{SR}^t$ *is a map*

$$\sigma_{\mathrm{SR}^t} : S_{\mathrm{SR}} \times \Phi \to A_{\mathrm{SR}}.$$

*Then* $v^*_{\mathrm{LR}} \leq v^*_Q$.

*Proof in Appendix A.6.*

Theorem 4 states that whether the long-lived player benefits from being committed to a $Q$-learning algorithm depends on the short-run players' information about past interactions. When short-run players observe the outcome of past interactions, there exists an equilibrium in which the strategic long-run achieves a higher payoff than a $Q$-learning algorithm can attain for any parameters $\langle Q^0, (\alpha^t), (\varepsilon^t)\rangle$. In contrast, when the short-run players do not observe the outcome of past interactions, there exist parameters of the algorithm such that the $Q$-learning algorithm achieves a higher payoff than the strategic long-run player in any equilibrium. In both cases, the gap is strict for some games.[39]

To illustrate the intuition behind the result, suppose that $|\Phi| = |S_Q| = 1$. When short-run players observe past outcomes they can provide incentives to the strategic long-run player by threatening to punish him if he deviates. The algorithm, however, does not take the possibility of punishment into account. When the short-run players' strategy is constant across periods, the algorithm will learn to play a (myopic) best-response to it. The algorithm fails to anticipate that the short-run players change their behavior as a response to it's learning.

When the short-run players do not observe past outcomes, they cannot condition their strategy on the long-run player's actions and thus cannot provide intertemporal incentives to him. Consequently, the long-run player must play a myopic best-response to the short-run players' strategy in almost all periods. Since the short-run players are in best-reply, the strategic player's payoff must equal an equilibrium payoff of the static game.

The algorithm, however, need not play a myopic best-response to $\sigma_{\mathrm{SR}^t}$ in any period

holds: the supremum of all $\epsilon$-equilibrium payoffs of this auxiliary game (a Nash equilibrium need not exist) for the algorithmic player equals $v^*_Q$.

[39] See Examples 3 and 4 in Appendix A.6.

$t$, even absent experimentation. The reason is that the $Q$-learning algorithm's play exhibits some inertia due to its slow learning. When the short-run players play a strategy independent of the period in which they are active, the algorithm will eventually converge to playing a best-response. However, this can induce the short-run players to alter their action, changing in turn the algorithmic player's best-response. As a consequence, play need not correspond to a Nash equilibrium of the static game. The algorithmic player can thus achieve a payoff strictly above any Nash equilibrium payoff of the static game. In contrast, a strategic long-run player knows the action $\mathrm{SR}^t$ plays in equilibrium, and hence can play a (myopic) best-response. It is the $Q$-learning algorithm's failure to play a best-response in every period that leads to higher payoffs than a strategic player can obtain.

We remark that the algorithm does not achieve higher payoffs than the strategic player simply because it has commitment power. Indeed, the long-run player's commitment to the algorithm does not change depending on whether the short-run players observe past outcomes or not. Yet, the comparison between the payoffs attained by the algorithm and by the strategic player depend on the short-run players' information about past outcomes. The role commitment to the algorithm plays is more subtle. Under complete information about the payoff functions, the long-run player would like to commit to the Stackelberg action. Commitment to the Stackelberg action is beneficial when the Stackelberg outcome is not a Nash equilibrium of the auxiliary simultaneous-move game so that the Stackelberg action is not a (myopic) best-response to the short-run player's action. Behaving according to a $Q$-learning algorithm provides a different kind of commitment: the algorithm is committed to learning and playing a best-response to any stationary strategy of the short-run players. Indeed, if the short-run players choose a constant action for a long enough horizon, the algorithm will play a best-response to that action. Commitment to learning a best-response against every stationary strategy of the short-run players is not optimal: there exist games in which the long-run player receives a strictly higher payoff when committed to not playing a best-response against a stationary strategy of the short-run players.[40]

# 4. Related literature

The algorithm we consider, $Q$-learning, was developed by Watkins (1989) and Watkins and Dayan (1992) for single-agent, Markov decision problems. However, $Q$-learning has been applied to study the interaction between multiple algorithms (so called multi-agent reinforcement learning); see Sandholm and Crites (1996), Leslie and Collins (2005), Rodrigues Gomes and Kowalczyk (2009), and Kianercy

[40] See the discussion of Examples 3 and 4.

and Galstyan (2012) to name but a few.[41] Bertrand et al. (2023), Dolgopolov (2024) and Schäfer (2022) study under what conditions two $Q$-learning algorithms play collusive outcomes in a Prisoner's Dilemma. More recently, Banchio and Mantegazza (2023), Cartea et al. (2022), and Possnig (2024) analytically study long-run outcomes of multiple reinforcement learning interacting with each other.[42] Our contribution to that literature is two-fold. First, we consider the interaction between an algorithm and myopically best-responding players as opposed to the interaction between multiple algorithms. Second, we delineate what properties of $Q$-learning extend from a stationary single-player environment to a strategic multi-player environment.

$Q$-learning has become the workhorse model for learning algorithms in the economics literature, and has been applied to study several topics. The most prominent application is to algorithmic collusion. Early papers are Kephart, Hanson, and Greenwald (2000), Gerald Tesauro and Kephart (2002), and Waltman and Kaymak (2008). More recently, the topic has gained renewed interest in Calvano et al. (2020), Klein (2021) and Asker, Fershtman, and Pakes (2022).[43] These papers study oligopoly games played by multiple $Q$-learning algorithms. Wang et al. (2023) study a pricing game between an artificial intelligence – modelled as a $Q$-learning algorithm – and a firm using a heuristic pricing rule. Johnson, Rhodes, and Wildenbeest (2023) study the design of a platform's recommendation rule when sellers set prices using $Q$-learning algorithms. Johnson, Rhodes, and Wildenbeest (2024) examine algorithmic steering of consumers on platforms and the effects of advertisement. Barberis and Jin (2023) use $Q$-learning to model how boundedly-rational agents make investment decisions. Decarolis et al. (2023) study the effect of privacy restrictions on ad auctions when advertisers submit bids using $Q$-learning algorithms. Our paper focuses on the interaction between a single algorithm and a myopically best-responding player. We do not restrict the class of games we study, but allow for general finite games.[44]

Our second contribution is to the understanding of the interaction between

[41] There is a literature that aims to design variants of $Q$-learning that lead to equilibrium play when employed by all players; see, e.g., Arslan and Yüksel (2016).

[42] The approximation techniques in Cartea et al. (2022) and Possnig (2024) are not applicable in our setting because the short-run players' best-responses lead to a violation of the required Lipschitz continuity.

[43] See also Calvano et al. (2019, 2021), Asker, Fershtman, and Pakes (2023), Abada and Lambin (2023), and Qiu et al. (2023). Werner (2023) studies collusion with human and algorithmic price-setters in lab experiments. Hettich (2021) studies collusion of Deep $Q$-Network algorithms that combine $Q$-learning with function approximation via deep neural networks. Hansen, Misra, and Pai (2021) assume that firms employ upper-confidence bound algorithms instead of $Q$-learning algorithms. A critique of the literature can be found in Dorner (2021), Boer, Meylahn, and Schinkel (2022), and Lambin (2024).

[44] The aforementioned papers are based on simulations whereas our methods are analytical.

strategic players and algorithms in games. Waizmann (2024) studies the repeated play of a patient (instead of a myopic), strategic player against a $Q$-learning algorithm. That paper focuses on whether the algorithm can be manipulated as opposed to what the algorithm can learn. Deng, Schneider, and Sivan (2019), Mansour et al. (2022), and D'Andrea (2023) derive bounds on the payoffs a patient strategic player can achieve when interacting with no-regret algorithms.[45] Our paper focuses on whether the algorithm can learn the Stackelberg outcome in a strategic environment whereas those papers aim to characterize the strategic player's maximum payoff.[46]

Thirdly, we contribute to the literature on algorithms in economics. Salcedo (2015), Lamba and Zhuk (2022),[47] and Levine (2023)[48] study pricing games in which sellers choose algorithms. The algorithms they consider are finite automata. Sellers learn each others' algorithms and have the opportunity to adjust their own algorithms at random times. Brown and MacKay (2023b) study how an algorithm's speed of adjustment affects competition.[49] In contrast to our paper, these papers feature no learning dynamics: all players know the payoff functions. Consequently, the role algorithms play in these paper differ from ours. Rather than an instrument for learning, the algorithms in those papers serve as commitment devices.

Lastly, we contribute to the literature on learning in (repeated) games; see Fudenberg and Levine (1998), Young (2004), and Hart and Mas-Colell (2013). In contrast to this literature, we consider learning with two asymmetric players: one behaving according to a fixed reinforcement learning rule while the second one plays a myopic best-response.[50] Kalai and Lehrer (1993), Nachbar (1997), and Wiseman (2005) study learning in repeated games. They focus on patient, strategic players whereas we consider a reinforcement learning algorithm interacting with myopic players.

There is a large literature on learning in Stackelberg games, i.e., repeated leader-follower interactions. Our paper differs from this literature in two regards. First, we consider a fixed algorithm as opposed to examining how to design an algorithm with desirable properties. Second, this literature either assumes that the follower

[45] See Guruganesh et al. (2024) for a similar exercise in a repeated contracting environment.

[46] A main challenge in those papers is to find conditions under which the strategic player can achieve a payoff strictly above his Stackelberg payoff.

[47] See also their follow-up paper, Lamba and Zhuk (in progress).

[48] Levine (2023) considers the cases of "observable commitment" and "unobservable commitment" to an algorithm – essentially, if the other player observes the choice of algorithm or not. This is different from our distinction between transparent and opaque algorithms which captures the short-run players' information about the algorithm's inputs. In Levine (2023), the algorithms' inputs are the outcome of past interaction and are observed by all players.

[49] See also Brown and MacKay (2023a).

[50] Some papers in this literature, e.g., Börgers and Sarin (1997), consider reinforcement learning procedures.

perfectly observes the leader's (mixed) action,[51] or assumes that the follower behaves according to a fixed algorithm.[52] We assume that the follower – the short-run players in our model – plays a history dependent best-response without assuming that the leader's – the algorithm's – action is perfectly observed.

# 5. Conclusion

Despite their pervasive use, the impact of learning algorithms on firm-consumer relations is not well explored. This paper seeks to fill this gap by studying how learning algorithms interact with strategic consumers. We examine conditions under which the algorithm achieves the maximal payoff even in an environment where consumers best-respond and therefore adapt their behavior.

We highlight the role of consumers' information about the algorithm's inputs. We examine transparent algorithms, whose inputs are observed by consumers, and opaque algorithms, whose inputs are hidden from consumers. We show that an algorithm performs better when consumers have less information about its inputs, and provide conditions such that consumer surplus is higher in that case as well. Moreover, we find that an algorithm can achieve higher payoffs than a strategic firm when consumers do not observe its inputs. Our results thus provide a novel rationale for why algorithms are often opaque.

# A. Proofs

## A.1. Proof of Theorem 1

*Proof.* When signalling is perfect, the short-run player assigns probability 1 to the algorithm having played $a_Q$ when observing the signal $\phi_{a_Q}$. Hence, when observing $\phi_{a_Q}$ the short-run players play $\mathrm{BR}(a_Q, s')$ in state $s' \in S_{\mathrm{SR}}$. Moreover, when the algorithm chooses $a_Q$ the signal $\phi_{a_Q}$ realizes with probability 1.

Consequently, when playing $a_Q$ in state $s \in S_Q$, the algorithm's (random) payoff is

$$u_Q(a_Q, \mathrm{BR}(a_Q, s), \omega).$$

This payoff is distributed according to $q(\cdot|s)$. Since the distribution of payoffs is sub-Gaussian, the expectation with respect to $q(\cdot|s)$ is well defined, and the

[51] See, e.g., Letchford, Conitzer, and Munagala (2009), Brückner and Scheffer (2011), Marecki, Gerry Tesauro, and Segal (2012), Balcan et al. (2015), Blum, Haghtalab, and Procaccia (2014), G. Yang, Poovendran, and Hespanha (2019), and Zhao et al. (2023).

[52] For example, Braverman et al. (2018), Fiez, Chasnov, and Ratliff (2019), Camara, Hartline, and Johnsen (2020), and Zrnic et al. (2021). Haghtalab, Podimata, and Yang (2023) assume the follower makes a calibrated forecast of the leader's action to which it then best-responds

variance of the random payoff is finite.

Hence, Theorem (Watkins) applies:

$$Q^t(a_Q, s) \to \mathbb{E}[u_Q(a_Q, \mathrm{BR}(a_Q, s), \omega)|s] \tag{2}$$

almost surely as $t \to \infty$. The limit on the right-hand side of equation 2 equals the expectation of the payoff the algorithm receives when playing $a_Q$ in state $s$. The maximum over the algorithm's actions of the quantity on the right-hand side equals the Stackelberg payoff $u_Q^{\mathrm{Stack}}(s)$.
Because actions are selected $1 - \varepsilon^t$-greedily and $\varepsilon^t \to 0$, the claim follows. □

## A.2. Proof of Theorem 2

*Proof.* By the hypothesis that $a^*(s)$ is a strict Nash equilibrium, there exists $\delta > 0$ such that

$$\mathbb{E}[u_Q(a_Q^*(s), a_{\mathrm{SR}}^*(s), \omega)|s] \geq \max_{a_Q \neq a_Q^*(s)} \mathbb{E}[u_Q(a_Q, a_{\mathrm{SR}}^*(s), \omega)|s] + 2\delta.$$

Choosing $K$ large enough, one may assume that $\varepsilon_K$ is close enough to 0 such that the best-response of $\mathrm{SR}^t$ in state $s$ is $a_{\mathrm{SR}}^*(s)$ irrespective of the realized signal when $Q^t(s, a_Q^*(s)) > Q^t(s, a_Q)$ for all $a_Q \neq a_Q^*(s)$ and $t \geq K$. Consequently, on the event $Q^t(s, a_Q^*(s)) > Q^t(s, a_Q)$, the payoff the algorithm receives at state $s$ when selecting action $a_Q$ is (the random variable) $u_Q(a_Q, a_{\mathrm{SR}}, \omega)_{|\omega \in s}$.
Applying Lemma 6 to each of the processes $Q^t(s, a_Q), a_Q \in A_Q$, the probability that $Q^t(s, a_Q^*(s)) < Q^t(s, a_Q)$ for some $a_Q \neq a_Q^*(s), t \geq K$ can be made smaller than $\xi$ by choosing $K$ large enough. □

## A.3. Proof of Theorem 3

*Proof.* First, the richness condition, Assumption 3 implies that for every $a_Q$ there exists $\omega \in \Omega$ such that

$$\min_{a_{\mathrm{SR}}} u_Q(a_Q, a_{\mathrm{SR}}, \omega) > \max_{a_{\mathrm{SR}}} u_Q(a'_Q, a_{\mathrm{SR}}, \omega) \ \forall a'_Q \neq a_Q.$$

Since the algorithm's information partition is rich, there exists $s_{a_Q} \in S_Q$ such that

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

Such a state $s_{a_Q}$ exists for each action $a_Q$ of the algorithm. Let $\underline{q} = \min_{a_Q} q(s_{a_Q})$. Since $q(\cdot)$ has full support and $S_Q$ is finite, $\underline{q} > 0$.

Second, let $\xi > 0$ be such that, for all $a_Q$, $\mathrm{BR}(a_Q)$ is the unique best-reply of the short-run players if the algorithmic player plays the action $a_Q$ with probability at least $\xi$. Such a $\xi$ exists because the short-run players' payoff function is generic. Suppose the signalling structure is $\gamma$-perfect. For each action $a_Q$, denote by $\phi_{a_Q}$ the signal that satisfies $p(\phi_{a_Q}|a_Q) \geq 1-\gamma$. Let $\gamma_1 > 0$ satisfy

$$\mathbb{P}^\tau[a_Q|\phi_{a_Q}] \geq \xi$$

for any fully mixed strategy $\tau \in \Delta(A_Q)$ with $\tau(a_Q) \geq \underline{q}/2$ for each $a_Q$; i.e., the posterior probability that the algorithmic player has played $a_Q$ after observing $\phi_{a_Q}$ is at least $\xi$ for any strategy that plays each action $a'_Q$ with probability at least $\underline{q}/2$. Such a $\gamma_1$ exists by Lemma 3.

Let $a_Q^{\mathrm{Stack}}(s)$ be the action that achieves the Stackelberg payoff in state $s$; that is,

$$a_Q^{\mathrm{Stack}}(s) = \arg\max_{a'_Q} \mathbb{E}[u_Q(a'_Q, \mathrm{BR}(a'_Q), \omega)|s].$$

Recall that Assumption 2 implies that the short-run players' best response against $a_Q$ does not depend on the state $s$. By genericity, the action $a_Q^{\mathrm{Stack}}(s)$ is unique for every state $s$. Let $\gamma_2$ be such that for each $s \in S_Q$,

$$\begin{aligned}&(1-\gamma)u_Q^{\mathrm{Stack}}(s) + \gamma \min_{a_{\mathrm{SR}}} \mathbb{E}[u_Q(a_Q^{\mathrm{Stack}}(s), a_{\mathrm{SR}}, \omega)|s] \\ &> (1-\gamma)\mathbb{E}[u_Q(a'_Q, \mathrm{BR}(a'_Q), \omega)|s] + \gamma \max_{a_{\mathrm{SR}}} \mathbb{E}[u_Q(a'_Q, a_{\mathrm{SR}}, \omega)|s] \ \forall a'_Q \neq a_Q^{\mathrm{Stack}}(s)\end{aligned}$$

for all $\gamma \leq \gamma_2$. Such a $\gamma_2$ exists by genericity of the algorithm's payoffs and the finiteness of $S_Q$. Choose $\overline{\gamma} = \min\{\gamma_1, \gamma_2\}$.

Third, fix a $\gamma$-perfect monitoring structure $(\Phi, p)$ with $0 < \gamma < \overline{\gamma}$. Let $(\sigma_{\mathrm{SR}^t})$ be an optimal joint strategy of the short-run players. Because of Assumption 3 and the hypothesis that the algorithm's information partition $S_Q$ is rich, Lemma 1 applies. Hence, there exists a period $T$ such that the short-run players believe the algorithm takes each action with probability at least $\overline{q}/2$ in each period $t \geq T$, irrespective of the experimentation rates $(\varepsilon^t)$.[53] By our choice of $\overline{\gamma}$, any optimal $\sigma_{\mathrm{SR}^t}$ must satisfy $\sigma_{\mathrm{SR}^t}(\phi_{a_Q}) = \mathrm{BR}(a_Q)$ for all $t \geq T$. Hence, Lemma 2 can be applied with $\Phi' = \{\phi_{a_Q} | a_Q \in A_Q\}$. Consequently, for every $\xi > 0$ chosen small enough, the

[53] The period $T$ depends on the experimentation probabilities $(\varepsilon^t)$.

$Q$-values are eventually almost surely contained in an interval. Specifically

$$\begin{aligned}
&\sum_{\phi_{a'_Q} \in \Phi'} p(\phi'_{a_Q}|a_Q)\mathbb{E}[u_Q(a_Q, \mathrm{BR}(a'_Q), \omega)|s] \\
&+ \left(1 - \sum_{\phi \in \Phi'} p(\phi|a_Q)\right) \min_{a_{\mathrm{SR}}} \mathbb{E}[u_Q(a_Q, a_{\mathrm{SR}}, \omega)|s] - \xi \\
&\leq Q^t(s, a_Q) \leq \\
&\sum_{\phi_{a'_Q} \in \Phi'} p(\phi'_{a_Q}|a_Q)\mathbb{E}[u_Q(a_Q, \mathrm{BR}(a'_Q), \omega)|s] \\
&+ \left(1 - \sum_{\phi \in \Phi'} p(\phi|a_Q)\right) \max_{a_{\mathrm{SR}}} \mathbb{E}[u_Q(a_Q, a_{\mathrm{SR}}, \omega)|s] + \xi,
\end{aligned}$$

Since $\gamma < \gamma_2$, the bounds are such that for each state $s$, the $Q$-value corresponding to the Stackelberg action is the highest $Q$-value in that state; that is, there exists a (random) period $\tilde{T}$ such that for all $t \geq \tilde{T}$ and states $s \in S_Q$,

$$Q^t(s, a_Q^{\mathrm{Stack}}(s)) > Q^t(s, a_Q) \ \forall a_Q \neq a_Q^{\mathrm{Stack}}(s).$$

Since the experimentation probabilities vanish, the algorithmic player chooses the Stackelberg action in each state $s$ with probability approaching 1. Moreover, the short-run players' strategies are such that they best-respond to the Stackelberg action, when played by the algorithm, with probability at least $1 - \gamma$. The claim then follows.

□

**Lemma 1.** *Suppose $s \in S_Q, a_Q \in A_Q$ are such that*

$$\min_{a_{\mathrm{SR}} \in A_{\mathrm{SR}}} \mathbb{E}\left[u_Q(a_Q, a_{\mathrm{SR}}, \omega) \mid s\right] > \max_{a_{\mathrm{SR}} \in A_{\mathrm{SR}}} \mathbb{E}\left[u_Q(a'_Q, a_{\mathrm{SR}}, \omega) \,\middle|\, s\right]$$

*for all $a'_Q \neq a_Q$.*
*Fix any joint strategy $(\sigma_{\mathrm{SR}^t})$ of the short-run players where*

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

*For every $\xi > 0$ there exists a (deterministic) period $T_\xi$ such that for all $t \geq T_\xi$,*

$$\mathbb{P}^{(\sigma_{\mathrm{SR}^t})}[Q^t(s, a_Q) > \max_{a'_Q \neq a_Q} Q^t(s, a'_Q)] \geq 1 - \xi.$$

*Proof.* Denote by $H^t$ the set of histories of length $t$, i.e.,

$$H^t = (S_Q \times A_Q \times U_Q)^t .$$

Denote the set of outcomes by

$$H^\infty = (S_Q \times A_Q \times U_Q)^\infty .$$

Let $\{\mathcal{F}_t\}_t$ be the filtration generated by the $H^t$. Recall that any $h^t \in H^t$ determines the $Q$–values at period $t$, i.e., $Q^t(h^t)$. To simplify notation, denote by $\mathbb{P}[\cdot] = \mathbb{P}^{(\sigma_{\text{SR}^t})}[\cdot]$ the probability measure induced by the algorithm's and the short–run players' strategies. Note that the SR-players' strategy is measurable with respect to the filtration $\{\mathcal{F}_t\}_t$.

Because the experimentation probabilities $(\varepsilon^t)$ satisfy $\sum_t \varepsilon^t = \infty$ and the distribution $q$ on $\Omega$ has full support, the event $\{\omega^t \in s, a_Q^t = a_Q \quad \text{for infinitely many} \quad t\}$ occurs with $\mathbb{P}$-probability 1. Apply Lemma 5 to the processes $(Q^t(s, a_Q))_t, (Q^t(s, a'_Q)_t)_t$, and conclude that there exists a random time $\tilde{T}$ such that, for a $\eta > 0$ small enough,

$$Q^t(s, a_Q) \geq \inf_k \sum_{\phi \in \Phi} p(\phi | a_Q) \mathbb{E}\left[u_Q\left(a_Q, \sigma_{\text{SR}_k}(\phi), \omega^k\right) \mid \omega^k \in s\right] - \eta/2$$

$$Q^t(s, a'_Q) \leq \sup_k \sum_{\phi \in \Phi} p(\phi | a'_Q) \mathbb{E}\left[u_Q\left(a_Q\prime, \sigma_{\text{SR}_k}(\phi), \omega^k\right) \mid \omega^k \in s\right] + \eta/2$$

Moreover, $\tilde{T} < \infty$ $\mathbb{P}$-almost surely. We remark that the random time $\tilde{T}$ is a function of the outcome $h \in H^\infty$. $\tilde{T}$ is *not* a stopping time with respect to the filtration $\{\mathcal{F}_t\}_t$.
Choosing $\eta$ small enough, by the hypothesis that

$$\min_{a_{\text{SR}} \in A_{\text{SR}}} \mathbb{E}\left[u_Q(a_Q, a_{\text{SR}}, \omega) \mid s\right] > \max_{a_{\text{SR}} \in A_{\text{SR}}} \mathbb{E}\left[u_Q(a'_Q, a_{\text{SR}}, \omega) \;\middle|\; s\right]$$

for all $a'_Q \neq a_Q$, we conclude that for all $t \geq \tilde{T}$,

$$Q^t(s, a_Q) - \eta > Q^t(s, a'_Q) \quad \forall a_Q \neq a'_Q.$$

Since $\tilde{T} < \infty$ for $\mathbb{P}$-almost all $h \in H^\infty$, there exists a (deterministic) $T_\xi \in \mathbb{N}$ such that $\mathbb{P}[\tilde{T} \leq T_\xi] \geq 1 - \xi$. The claim follows. □

**Lemma 2.** *Let $\sigma_{\mathrm{SR}^t} : \Phi \to A_{\mathrm{SR}}$.*[54] *Suppose there exists $\Phi' \subset \Phi$ such that*

$$\sigma_{\mathrm{SR}^t}(\phi) = \sigma_{\mathrm{SR}_{t'}}(\phi) \equiv \sigma_{\mathrm{SR}}(\phi) \quad \forall t, t', \phi \in \Phi'.$$

*Fix $\xi > 0$. There exists a random time $T_\xi$ such that for every $s \in S_Q$ and $a_Q \in A_Q$,*

$$\sum_{\phi \in \Phi'} p(\phi|a_Q)\mathbb{E}[u_Q(a_Q, \sigma_{\mathrm{SR}}(\phi), \omega)|s] + \left(1 - \sum_{\phi \in \Phi'} p(\phi|a_Q)\right) \min_{a_{\mathrm{SR}}} \mathbb{E}[u_Q(a_Q, a_{\mathrm{SR}}, \omega)|s] - \xi$$
$$\leq Q^t(s, a_Q) \leq$$
$$\sum_{\phi \in \Phi'} p(\phi|a_Q)\mathbb{E}[u_Q(a_Q, \sigma_{\mathrm{SR}}(\phi), \omega)|s] + \left(1 - \sum_{\phi \in \Phi'} p(\phi|a_Q)\right) \max_{a_{\mathrm{SR}}} \mathbb{E}[u_Q(a_Q, a_{\mathrm{SR}}, \omega)|s] + \xi,$$

*for all $t \geq T_\xi$. Moreover, $T_\xi < \infty$ almost surely.*

*Proof.* For $a_Q \in A_Q, s \in S_Q$ fixed, denote by $\mathbb{1}\{(a_Q^t, s^t) = (a_Q, s)\}$ the random indicator variable that the state-action combination in period $t$ was $(a_Q, s)$. Let $n_t(s, a_Q)$ be the number of times the state-action pair $(a_Q, s)$ has been visited before time $t$.

The $Q$-values evolve as follows:

$$Q^{t+1}(s, a_Q) = Q^t(s, a_Q) + \alpha^{n_t(s,a_Q)}\mathbb{1}\{(a_Q^t, s^t) = (a_Q, s)\}\left(-Q^t(s, a_Q) + u_Q^t\right),$$

where $u_Q^t$ is the payoff the algorithm receives in period $t$. Conditional on $(s, a_Q)$ the payoff $u_Q^t$ is the random variable

$$u_Q^t = u_Q(a_Q, \sigma_{\mathrm{SR}^t}(\phi^t), \omega^t)$$

where $\phi^t$ is distributed according to $p(\cdot|a_Q)$ and $\omega^t$ according to $q_{\omega \in s}(\cdot)$. Note that $\phi^t$ and $\omega^t$ are independent, conditional on $(s, a_Q)$.
Let $\{\mathcal{F}_t\}$ be the $\sigma$-algebra generated by $(S_Q \times A_Q \times U_Q)^t$, and $\{\mathcal{G}_t\}$ the $\sigma$-algebra generated by $(S_Q \times A_Q \times U_Q)^t \times S_Q \times A_Q$.

Note that $\mathbb{1}\{(a_Q^t, s^t) = (a_Q, s)\}$ and $\alpha^{n_t(s,a_Q)}$ are $\mathcal{G}_t$-measurable. Define an alternative process $\tilde{Q}^t(s, a_Q)$ recursively by

$$\tilde{Q}^{t+1}(s, a_Q) = \tilde{Q}^t(s, a_Q) + \alpha^{n_t(s,a_Q)}\mathbb{1}\{(a_Q^t, s^t) = (a_Q, s)\}\left(-\tilde{Q}^t(s, a_Q) + u_Q^t - \mathbb{E}[u_Q^t|\mathcal{G}_t]\right),$$

and $\tilde{Q}^0(s, a_Q) = Q^0(s, a_Q)$.

[54] Extending the Lemma to allow for mixed strategies of the short-run players is straightforward.

The event $\{\mathbb{1}_{(a_Q^t,s^t)=(a_Q,s)} \text{ for infinitely many } t\}$ occurs with probability 1 because the experimentation probabilities $(\varepsilon^t)_t$ satisfy $\sum_t \varepsilon^t = \infty$. Consequently, $\tilde{Q}^t(s,a_Q) \to 0$ almost surely. One computes that, almost surely,

$$\begin{aligned}\mathbb{E}[u_Q^t|\mathcal{G}_t] &= \sum_{\phi\in\Phi} p(\phi|a_Q)\mathbb{E}[u_Q(a_Q,\sigma_{\mathrm{SR}^t}(\phi),\omega)|s_Q] \\ &\geq \sum_{\phi\in\Phi'} p(\phi|a_Q)\mathbb{E}[u_Q(a_Q,\sigma_{\mathrm{SR}}(\phi),\omega)|s_Q] + \left(1-\sum_{\phi\in\Phi'}\right)\min_{a_{\mathrm{SR}}}\mathbb{E}[u_Q(a_Q,a_{\mathrm{SR}},\omega)|s_Q].\end{aligned}$$

The first line uses that the distribution of the signal $\phi$ is independent of $\omega$, conditional on $s_Q$ and $a_Q$. The inequality follows from our hypothesis on the strategy $\sigma_{\mathrm{SR}^t}$ for signals $\phi \in \Phi'$. Note that the second term on the last line does not depend on the period $t$.

By an analogous computation,

$$\begin{aligned}\mathbb{E}[u_Q^t|\mathcal{G}_t] \leq &\sum_{\phi\in\Phi'} p(\phi|a_Q)\mathbb{E}[u_Q(a_Q,\sigma_{\mathrm{SR}}(\phi),\omega)|s_Q] \\ &+ \left(1-\sum_{\phi\in\Phi'}\right)\max_{a_{\mathrm{SR}}}\mathbb{E}[u_Q(a_Q,a_{\mathrm{SR}},\omega)|s_Q].\end{aligned}$$

Applying Lemma 5 to each process $Q^t(s,a_Q)$, the claim follows. □

**Lemma 3.** *Suppose Assumption 2 hold. Let $a_Q, \phi_{a_Q}$ be as in definition 7. For any $\underline{\mu} > 0$, there exists $\overline{\gamma} > 0$ such that*

1. *if $\mathbb{P}[a_Q^t = a_Q] \geq \underline{\mu}$ for all $a_Q \in A_Q$,*
2. *and $(\Phi, p(\cdot|\cdot))$ is $\gamma$-perfect for $\gamma \leq \overline{\gamma}$,*

*then $\mathrm{SR}^t$ plays $\mathrm{BR}(a_Q)$ after observing signal $\phi_{a_Q}$, i.e., $\sigma_{\mathrm{SR}^t}(\phi_{a_Q}) = \mathrm{BR}(a_Q)$.*[55]

*Proof.* Because the short-run players' payoffs are generic, there exists $0 < \eta < 1$ such that $\mathrm{SR}^t$ plays $\mathrm{BR}(a_Q)$ if the probability that $a_Q^t = a_Q$ is at least $\eta$. Since $A_Q$ is finite, $\eta$ can be taken independent of $a_Q$.
Since the signal distribution $p$ has full support, one computes that

$$\mathbb{P}[a_Q^t = a_Q|\phi_{a_Q}] \geq \frac{(1-\gamma)\underline{\mu}}{(1-\gamma)\underline{\mu} + \gamma(1-\underline{\mu})}.$$

The RHS of the inequality is greater than $\eta$ for any $\underline{\mu} > 0$ if $\gamma$ is close enough to 0. □

[55] This Lemma is reminiscent of Lemma 1 in Van Damme and Hurkens (1997).

## A.4. Proof of Proposition 1

*Proof.* 1. Theorem 3 implies that

$$W^{Q,\text{opa}}_{(\Phi,p_n)} \to \mathbb{E}[u^{\text{Stack}}_Q(s)].$$

which shows the first assertion.
Let $M$ be a bound on the algorithmic player's expected payoff. When the algorithm is transparent, the expected payoff the algorithm receives in period $t$ and state $s$ is bounded from above by $(1-\varepsilon^t)u^{\text{Stack}}_Q(s) + \varepsilon^t M$ for all $t$ large enough. Hence, $\limsup W^{Q,\text{tra}}_{(\Phi,p_n)}$ is bounded from above by $\mathbb{E}[u^{\text{Stack}}_Q(s)]$.

2. When the short-run player does not have a dominant action, there exists $a'_Q \neq a^\dagger_Q$ such that

$$\text{BR}(a'_Q) \neq \text{BR}(a^\dagger_Q).$$

By Assumption 3, there exists $\omega \in \Omega$ such that

a) $u_Q(a^\dagger_Q, \text{BR}(a^\dagger_Q), \omega) > u_Q(a_Q, \text{BR}(a_Q), \omega)\ \forall a_Q \neq a^\dagger_Q$;

b) $u_Q(a'_Q, \text{BR}(a^\dagger_Q, \omega) > u_Q(a^\dagger_Q, \text{BR}(a^\dagger_Q), \omega)$;

c) $u_Q(a'_Q, \text{BR}(a'_Q), \omega) > u_Q(a_Q, \text{BR}(a'_Q), \omega)\ \forall a_Q \neq a'_Q$

d) $u_Q(a'_Q, \text{BR}(a'_Q), \omega) > u_Q(a, \omega)$
$\forall a \in A_Q \times A_{\text{SR}} \setminus \{(a'_Q, \text{BR}(a'_Q)), (a'_Q, \text{BR}(a^\dagger_Q), (a^\dagger_Q, \text{BR}(a^\dagger_Q))\}$.

The first condition states that $a^\dagger_Q$ is the Stackelberg action in the auxiliary game $G(\{\omega\})$. The second condition states that this Stackelberg outcome is not a Nash equilibrium of the auxiliary game $G(\{\omega\})$. The third condition states that there is a Nash equilibrium of $G(\{\omega\})$. The last condition requires that the only action pairs that yield a higher payoff for the algorithmic player than the Nash equilibrium $(a'_Q, \text{BR}(a'_Q))$ are the Stackelberg outcome and $(a'_Q, \text{BR}(a^\dagger_Q)$.
By the hypothesis that $S_Q$ is a rich partition, there exists $s \in S_Q$ such that $\mathbb{E}[u_Q(\cdot, \omega')|s]$ induces the same preference relation over $A_Q \times A_{\text{SR}}$ as $u_Q(\cdot, \omega)$. In particular, the conditions on $a^\dagger_Q, a'_Q$ hold for the expected payoff conditional on $s$. Note that $a^\dagger_Q = a^{\text{Stack}}_Q(s)$.
When the algorithm is transparent, we know from the proof of Theorem 2 that, since the signalling structure has full support, there exists a period $K$

such that for all periods $t \geq K$

$$\sigma_{\mathrm{SR}^t}(h^{t-1}, s, \phi^t) = \mathrm{BR}(\arg\max_{a_Q} Q^t(h^{t-1})(s, a_Q)$$

for all signals $\phi$. That is, the short-run player $\mathrm{SR}^t$ plays a best-response to the greedy action in state $s$ irrespective of the signal $\phi^t$.

Assume toward a contradiction that the Stackelberg outcome is played in almost every period, i.e., the fraction of periods in which the Stackelberg outcome in state $s$ is played converges to 1 almost surely. Then it must be that for $t$ large enough, $u_Q^{\mathrm{Stack}}(s) \approx Q^t(s, a_Q^\dagger) > Q^t(s, a_Q)$ for all $a_Q \neq a_Q^\dagger$ and $\mathrm{SR}^t$ plays $\mathrm{BR}(a_Q^t)$. By condition 2, there exists a (random but almost surely finite) period $T$ such that $Q^t(s, a_Q') > Q^t(s, a_Q)$ for all $a_Q \neq a_Q'$.

Because $(a_Q', \mathrm{BR}(a_Q'))$ is a strict Nash equilibrium of the auxiliary game $G(s)$, the algorithmic player receives a strictly lower payoff when playing $\tilde{a}_Q \neq a_Q'$ against $\mathrm{BR}(a_Q')$ Hence, there exists a period $N$, a sequence of actions of the algorithm $(a_Q^t)_{t=T}^N$ and realizations $(\omega^t)_{t=T}^N, \omega^t \in s$ such that that if $(a_Q^t, \mathrm{BR}(a_Q'))$ is played and $\omega^t$ realized in periods $t = T, \ldots, N$, (i) $Q_N(s, a_Q')$ is in a neighborhood of radius $\delta/2$ of $\mathbb{E}[u_Q(a_Q', \mathrm{BR}(a_Q'), \omega)|s]$ for a small $\delta > 0$, (ii) $\mathbb{E}[u_Q(a_Q', \mathrm{BR}(a_Q'), \omega)|s] - \delta > Q_N(s, a_Q)$ for all $a_Q \neq a_Q'$, and (iii) $Q^t(s, a_Q') > Q^t(s, a_Q)$ for all $a_Q \neq a_Q'$ and $t = T, \ldots, N$. Condition (iii) ensures that $\mathrm{SR}^t$ plays $\mathrm{BR}(a_Q')$ in periods $t = T, \ldots, N$. Because $\varepsilon^t > 0$, such a sequence of actions occurs with positive probability. However, by an argument similar to the proof of Theorem 2, conditional on the event (i) and (ii),

$$\mathbb{P}\left[\lim_{T\to\infty} \frac{\sum_{t=N}^{N+T} \mathbb{1}\{(a_Q^t, a_{SR}^t, s^t) = (a_Q', \mathrm{BR}(a_Q'), s)\}}{\sum_{t=N}^{N+T} \mathbb{1}\{s^t = s\}} = 1\right] \geq \xi$$

for a $\xi > 0$. Consequently,

$$\lim_{T\to\infty} \frac{\sum_{t=0}^{T} u_Q(a_Q^t, a_{\mathrm{SR}}^t, \omega^t)\mathbb{1}\{\omega^t \in s\}}{\sum_{t=0}^{T} \mathbb{1}\{\omega^t \in s\}} < u_Q^{\mathrm{Stack}}(s)$$

with positive probability, a contradiction. Consequently, $W_{(\Phi,p)}^{Q,\mathrm{tra}} < \mathbb{E}[u_Q^{\mathrm{Stack}}(s)]$.

3. For each $s$, let $\underline{w}(s)$ equal the lowest payoff for the algorithmic player in any strict Nash equilibrium of the auxiliary game $G(s)$ if $G(s)$ admits a strict Nash equilibrium, and $\underline{w}(s) = u_Q^{\mathrm{Stack}}(s)$ for all other states. By Assumption 3 and the hypothesis that $S_Q$ is a rich partition, $\mathbb{E}[\underline{w}(s)] < \mathbb{E}\left[u_Q^{\mathrm{Stack}}(s)\right]$. Let $\delta = \mathbb{E}\left[u_Q^{\mathrm{Stack}}(s)\right] - \mathbb{E}[\underline{w}(s)]$. By Theorem 3, $W_{(\Phi,p_n)}^{Q,\mathrm{opa}} \geq \mathbb{E}\left[u_Q^{\mathrm{Stack}}(s)\right] - \delta/3$ for all

$n$ large enough. Moreover, for all $n$ large enough, Theorem 2 and its corollary there exists parameters of the algorithm such that $W^{Q,\text{tra}}_{(\Phi,p_n)} \leq \mathbb{E}[\underline{w}(s)] + \delta/3$. The claim follows.

□

## A.5. Proof of Proposition 2

*Proof.* 1. By Lemma 4, in a state $s$ such that for some $a_Q$ equation (1) holds, the algorithm plays $a_Q$ eventually (up to experimentation), irrespective of the short-run players' strategy. In particular, this holds when the algorithm is transparent and when the algorithm is opaque. Consequently, the highest asymptotic payoff the short-run player obtain in such a state is $\mathbb{E}\left[u_{\text{SR}}(a_Q, \text{BR}(a_Q), \omega)\right]$.

By Theorem 3 and Remark 1, the opaque algorithm learns the Stackelberg action when signalling is precise enough. Consequently, when the algorithm is opaque the expected payoff the short-run player receives for all periods $t$ large enough converges to $\mathbb{E}[u_{\text{SR}}(a_Q^{\text{Stack}}(s), \text{BR}(a_Q^{\text{Stack}}(s), \omega)]$ as $n \to \infty$. Given the hypotheses of the claim, this is the highest expected payoff the short-run players can achieve.

2. The proof follows along the same lines as the third part in the proof of Proposition 1.

□

**Lemma 4.** *Suppose that for some state $s \in S_Q$ and some action $a_Q \in A_Q$ equation (1) holds, i.e.,*

$$\min_{a_{\text{SR}} \in A_{\text{SR}}} \mathbb{E}\left[u_Q(a_Q, a_{\text{SR}}, \omega)|s\right] > \max_{a_{\text{SR}} \in A_{\text{SR}}} \mathbb{E}\left[u_Q(a'_Q, a_{\text{SR}}, \omega)|s\right]$$

*for all $a'_Q \neq a_Q$.*

*Fix any joint strategy $(\sigma_{\text{SR}^t})$ of the short-run players where*

$$\sigma_{\text{SR}^t} : (S_Q \times A_Q \times U_Q)^{t-1} \times S_Q \times \Phi \to A_{\text{SR}}.$$

*Then*

$$\left\{h \in (S_Q \times A_Q \times U_Q)^\infty \;\middle|\; \exists T \in \mathbb{N} : Q^t(h)(s, a_Q) > Q^t(h)(s, a'_Q)\ \forall t \geq T\right\}$$

*occurs with probability* 1.

*Proof.* Let $s, a_Q$ satisfy equation (1). Let $a'_Q \neq a_Q$. Consider the stochastic process $(Q^t(s, a_Q))_t$. Since $(s^t, a_Q^t) = (s, a_Q)$ for infinitely many $t$ and $Q^{t+1}(s, a_Q) =$

$Q^t(s, a_Q)$ if $(s^t, a_Q^t) \neq (s, a_Q)$, one may assume that $(s^t, a_Q^t) = (s, a_Q)$ for all $t$. Letting $\{\mathcal{F}_t\}_t$ be the filtration generated by the SR players' information, $v^t_{a_{\text{SR}}}(\omega) = u_Q(a_Q, a_{\text{SR}}, \omega)$ and $\xi^t_{a_{SR}}(h^t) = \sigma_{\text{SR}}(h^t)(a_{SR})$ the SR player's behavioral strategy at $h_t$, one can apply Lemma 5 to the processes $(Q^t(s, a_Q))$. One concludes that for every $\delta > 0$ there exists $T(\delta, a_Q)$ almost surely finite such that

$$Q^t(s, a_Q) \geq \min_{a_{\text{SR}} \in A_{\text{SR}}} \mathbb{E}[u_Q(a_Q, a_{\text{SR}}, \omega)|s] - \delta$$

for all $t \geq T(\delta, a_Q)$.
By an analogous reasoning, once concludes that
$Q^t(s, a'_Q) \leq \max_{a_{\text{SR}} \in A_{\text{SR}}} \mathbb{E}[u_Q(a'_Q, a_{\text{SR}}, \omega)|s] + \delta$ for all $t \geq T(\delta, a'_Q)$ for an almost surely finite $T(\delta, a'_Q)$. Choosing $\delta > 0$ such that

$$\min_{a_{\text{SR}} \in A_{\text{SR}}} \mathbb{E}[u_Q(a_Q, a_{\text{SR}}, \omega)|s] - \max_{a_{\text{SR}} \in A_{\text{SR}}} \mathbb{E}[u_Q(a'_Q, a_{\text{SR}}, \omega)|s] > 2\delta,$$

the claim follows. □

## A.6. Proof of Theorem 4

*Proof.* First, note by Assumptions 1 and 2, it is without loss of generality to assume that $|S_{\text{SR}}| = 1$; otherwise, all statements hold conditional on each $s \in S_{\text{SR}}$.

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 $\phi$ and takes an action $a_{\text{SR}} \in A_{\text{SR}}$. For each $a \in A_Q \times A_{\text{SR}}$ the payoffs of the short-run player is given by $\mathbb{E}[u_{\text{SR}}(a, \omega)]$ 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 $G(S_Q, (\Phi, p))$.[56] To ensure existence of a Nash equilibrium of the repeated game between the strategic LR player and the short-run players in pure strategies, we make the following assumption.

**Assumption 4.** *The game $G(S_Q, (\Phi, p))$ admits a Nash equilibrium in pure strategies.*

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

Moreover, assume payoffs in (normal form game of) $G(S_Q, (\Phi, p))$ are generic.This is a slightly stronger assumption than the genericity assumption on the payoff functions we impose throughout; see Section 2. The reason is that the genericity condition here also includes condition on the distributions $q$ and $p$.

We divide the proof into two cases. In the first case, we characterize payoffs when the short-run players observe the outcome of past interactions. In the second case, we characterize payoffs when the short-run players do not observe the outcome of past interactions.

**Case 1: suppose** $\sigma_{\mathrm{SR}^t} : (S_Q \times A_Q \times U_Q)^{t-1} \times \Phi \to A_{\mathrm{SR}}$**.** For $\tau_{\mathrm{LR}} : S_Q \to A_Q$, call $\tau_{\mathrm{SR}} : \Phi \to \Delta(A_{\mathrm{SR}})$ a best-response (in the auxiliary game $G(S_Q, (\Phi, p))$) to $\tau_{\mathrm{LR}}$ if

$$\begin{aligned}&\mathbb{E}^{(\tau_{\mathrm{LR}}, \tau_{\mathrm{SR}})}\left[u_{\mathrm{SR}}\left(a_{\mathrm{LR}}, a_{\mathrm{SR}}(\phi), \omega\right)\right] \\ &\geq \mathbb{E}^{\left(\tau_{\mathrm{LR}}, \tau'_{\mathrm{SR}}\right)}\left[u_{\mathrm{SR}}\left(a_{\mathrm{LR}}, a_{\mathrm{SR}}(\phi), \omega\right)\right] \quad \forall \tau'_{\mathrm{SR}}.\end{aligned}$$

Denote the best-response to $\tau_{\mathrm{LR}}$ by $\mathrm{BR}(\tau_{\mathrm{LR}})$ Let

$$v^* = \sup_{\tau_{\mathrm{LR}} : S_Q \to A_{\mathrm{LR}}} \mathbb{E}^{(\tau_{\mathrm{LR}}, \mathrm{BR}(\tau_{\mathrm{LR}}))}\left[u_{\mathrm{LR}}(a, \omega)\right],$$

i.e., the highest payoff the long-lived player can achieve in auxiliary game $G(S_Q, (\Phi, p))$ by committing to a pure strategy and the short-run player playing a best-response. Let $\tau^*_{\mathrm{LR}}$ be the strategy of the long-lived player that achieves this maximum payoff and $\tau^*_{\mathrm{SR}} = \mathrm{BR}(\tau^*_{\mathrm{LR}})$.

Construct an equilibrium $(\sigma_{\mathrm{LR}}, (\sigma_{\mathrm{SR}^t})$ of the repeated between the strategic long-run player LR and the short-run players as follows. Fix a pure Nash equilibrium $(\tau_{\mathrm{LR}}, \tau_{\mathrm{SR}})$ of the auxiliary game $G(S_Q, (\Phi, p)))$. Let

$$R^t = \left\{r^t \in (S_Q \times A_Q) \,|\, r^t = (s^1, \tau_{\mathrm{LR}}(s^1), \ldots, s^t, \tau_{\mathrm{LR}}(s^t))\right\}.$$

$$\sigma_{\mathrm{LR}}(h^{t-1}, s^t) = \begin{cases} \tau^*_{\mathrm{LR}}(s^t) & t = 0 \vee h^{t-1} \in R^{t-1}; \\ \tau_{\mathrm{LR}}(s^t) & t \geq 1 \wedge h^{t-1} \notin R^{t-1}. \end{cases}$$

$$\sigma_{\mathrm{SR}^t}(h^{t-1}, \phi^t) = \begin{cases} \tau^*_{\mathrm{SR}}(\phi^t) & h^{t-1} \in R^{t-1}; \\ \tau_{\mathrm{SR}}(\phi^t) & h^{t-1} \notin R^{t-1}. \end{cases}$$

It is easy to see that $(\sigma_{\mathrm{LR}}, (\sigma_{\mathrm{SR}^t})$ thus defined constitute a Nash equilibrium with payoff $v^*$ for LR. Consequently, $v^*_{\mathrm{LR}} \geq v^*$.

Now consider the $Q$-learning algorithm for fixed parameters $\langle Q^0, (\alpha^t), (\varepsilon^t)\rangle$. Because $SR^t$ observes the history $h^{t-1} \in (S_Q \times A_Q \times U_Q)^{t-1}$, $\mathrm{SR}^t$ knows the $Q$-

values. Moreover, since $\varepsilon^t \to 0$, there exists $T \in \mathbb{N}$ such that for all $t \geq T$, $\text{SR}^t$ plays a best-response (in $G(S_Q, (\Phi, p))$) to

$$s \mapsto \arg\max_{a_Q \in A_Q} Q^t(s, a_Q)(h^{t-1})$$

at the history $h^{t-1}$. Consequently, the algorithm's expected payoff is less than $(1-\varepsilon^t)v^* + \varepsilon^t M$ for an upper bound $M$ on $\mathbb{E}[u_Q(\cdot,\cdot,\omega)]$. Consequently, $v_Q^* \leq v^*$.

**Case 2: suppose $\sigma_{\text{SR}^t} : \Phi \to A_{\text{SR}}$.** Let $(\tau_{\text{LR}}, \tau_{\text{SR}})$ be the pure Nash equilibrium of the auxiliary game $G(S_Q, (\Phi, p))$ that maximizes the algorithmic player's payoff. Denote by $\overline{v}$ the expected payoff of the long-lived player from $(\tau_{\text{LR}}, \tau_{\text{SR}})$. We claim: $v_{\text{LR}}^* \leq \overline{v}$.

Let $\sigma = (\sigma_{\text{LR}}, (\sigma_{\text{SR}^t}))$ be an equilibrium of the repeated game. Denote by $v_{\text{LR}}^t = \mathbb{E}^\sigma[u_Q(a_{\text{LR}}^t, a_{\text{SR}}^t, \omega^t)]$ the expected payoff of the long-run player in period $t$. Suppose that $\lim_{T\to\infty} \sum_{t=0}^{T-1} v_{\text{LR}}^t/T$ exists.[57]

Denote by $\tau_{\text{LR}}^t$ the (mixed) strategy in the auxiliary game $G(S_Q, (\Phi, p))$ defined by

$$\tau_{\text{LR}}^t(s) = \sum_{h^{t-1} \in \left(S_Q \times A_Q \times U_Q\right)^{t-1}} \mathbb{P}^\sigma[h^{t-1}]\sigma_{\text{LR}}(h^{t-1}, s).$$

Since $\sigma$ is an equilibrium, $\sigma_{\text{SR}^t} = \text{BR}(\tau_{\text{LR}}^t)$.

Fix a $\delta > 0$ and let

$$\mathbb{N} \supset \Gamma(\delta) = \{t | v_{\text{LR}}^t \geq \overline{v} + \delta\}.$$

Suppose toward a contradiction that $\Gamma(\delta)$ has positive density, i.e., $\sum_{t=0}^{T-1} \frac{\mathbb{1}\{t\in\Gamma(\delta)\}}{T} = \mu > 0$. Since $v_{\text{LR}}^t \geq \overline{v} + \delta$ and $\sigma_{\text{SR}^t} = \text{BR}(\tau_{\text{LR}}^t)$, it cannot be that $\tau_{\text{LR}^t} = \text{BR}_{\text{LR}}(\sigma_{SR^t})$. Moreover, since $\delta > 0$ there exists $\delta_1 > 0$ such that $\tau_{\text{LR}^t}$ is not within $\delta_1$ of $\text{BR}_{\text{LR}}(\sigma_{SR^t})$ for any $t \in \Gamma(\delta)$. Hence, there is $\delta_3 > 0$ such that LR can raise his payoff by $\delta_3$ by playing $\sigma'_{LR}(h^{t-1}) = \text{BR}_{\text{LR}}(\sigma_{\text{SR}^t})$ instead of $\sigma_{LR}(h^{t-1})$ for $t \in \Gamma(\delta)$, contradicting the hypothesis that $\sigma$ is an equilibrium. Thus, for every $\delta > 0$, $\Gamma(\delta)$ has density 0. Consequently, $\lim_{T\to\infty} \sum_{t=0}^{T-1} v_{\text{LR}}^t/T \leq \overline{v}$ so that $v_{\text{LR}}^* \leq \overline{v}$.

Now consider the $Q$-learning algorithm. Let $(\tau_{\text{LR}}, \tau_{\text{SR}})$ be the pure Nash equilibrium of the auxiliary game $G(S_Q, (\Phi, p))$ that maximizes the algorithmic player's payoff. Denote by $\overline{v}$ the expected payoff of the long-lived player from $(\tau_{\text{LR}}, \tau_{\text{SR}})$. Applying Claim 1 for trivial partitions, i.e., $|P^t| = 1$ for each $t$, there exists a sequence $\langle Q_n^0, (\alpha^t)_n, (\varepsilon^t)_n\rangle_n$ of parameters for the algorithm such that the algorithm's payoff with parameters $\langle Q_n^0, (\alpha^t)_n, (\varepsilon^t)_n\rangle$ is at least $\overline{v} - 1/n$. Consequently, $v_Q^* \geq \overline{v}$. □

The next claim provides a partial characterization of the payoffs *some* algorithm

[57] Adjusting the proof if the limit does not exist is straightforward.

can achieve when the short-run players can condition their actions on a subset of the outcomes of past interactions. For any period $t$ let $P^t$ be a partition of the set of histories[58]

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

A strategy for the short-run player in period $t$, $\mathrm{SR}^t$ is a map

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

**Claim 1.** *Fix any sequence of partitions $\{P^t\}$.*
*For any $\xi > 0$ there exists an open set of parameters $\langle Q^0, (\alpha^t), (\varepsilon^t)_t \rangle$ and for each $(s, a_Q)$ an open neighborhood $\mathcal{O}_{(s,a_Q)}$ of*

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

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

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

*in any equilibrium $\left(\sigma^*_{\mathrm{SR}^t}\right)$ where*

$$\sigma^*_{\mathrm{SR}^t} : P^{t-1} \times \Phi \to \Delta(A_{\mathrm{SR}})$$

*Proof.* Fix $\xi > 0$.

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

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

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

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

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

[58] When the partition $P^t$ is trivial in every period $t$, i.e., $|P^t| = 1$, the algorithm is opaque. When $P^t = \{(S_Q \times A_Q \times U_Q)^t\}$ for each $t$, the algorithm is transparent.

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 $\tau_{\text{SR}}$ 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 $\text{SR}^t$'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 $s, t$,

$$Q^t(s, \tau_Q(s)) > Q^t(s, a_Q)\ \forall a_Q \neq \tau_Q(s) \tag{3}$$

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

$$Q^{M_2}(s, a_Q) \in \mathcal{O}_{(s,a_Q)},$$

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,

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

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

**Example 3.** For this example, suppose that the signal about the long-run player's action is uninformative, i.e., $|\Phi| = 1$. Assume expected payoffs are given as in Figure 3. Suppose that the prior probability over the states satisfies $1/2 < q(s_1)$.

Suppose that $S_Q = \{s_1, s_2, s_3\}$ Suppose the short-run players observe past interactions so that the strategy for $\text{SR}^t$ is a map

$$\sigma_{\text{SR}^t} : (S_Q \times A_Q \times U_Q)^{t-1} \to \Delta(A_{\text{SR}}).$$

| | | SR | | | SR | | | SR | |
|---|---|---|---|---|---|---|---|---|---|
| | | $L$ | $R$ | | $L$ | $R$ | | $L$ | $R$ |
| long-lived player | $T$ | $2,1$ | $1,0$ | $T$ | $0,1$ | $0,0$ | $T$ | $0,1$ | $20,0$ |
| | $B$ | $0,0$ | $0,1$ | $B$ | $1,0$ | $5,1$ | $B$ | $1,0$ | $1,1$ |
| | | $s_1$ | | | $s_2$ | | | $s_3$ | |

Figure 3: The game for Example 3.

| | | $SR$ | | |
|---|---|---|---|---|
| | | L | M | R |
| long-lived player | T | $2,3$ | $3,0$ | $-1,-1$ |
| | M | $4,0$ | $1,1$ | $-2,-2$ |
| | B | $-3,-3$ | $-4,-4$ | $0,0$ |

Figure 4: The payoff matrix for Example 4. Payoffs are deterministic.

The Stackelberg payoffs are $u^{\text{Stack}}(s_1) = 2$, $u^{\text{Stack}}(s_2) = 5$, and $u^{\text{Stack}}(s_3) = 1$. By Lemma 4, any $Q$-learning algorithm eventually plays $T$ in state $s_1$. The unique best-response by the short-run players is then to play $L$, irrespective of the action the algorithm takes in states $s_2$ and $s_3$. Consequently, eventually the algorithm plays $B$ in $s_2$ and $s_3$ to achieve an expected payoff of $1 + q(s_1)$.

Suppose the strategic long-run player LR plays $B$ in both states $s_1$ and $s_2$ and $T$ in state $s_3$. The best-response of the short-run player is then $R$. By the preceding proof, playing according to this strategy in every period is an equilibrium outcome of the repeated game. The expected payoff of the long-run player is $5q(s_2) + 20q(s_3)$. For an open set of parameters, this payoff is strictly greater than $2q(s_1) + 5q(s_2) + q(s_3)$, the expected Stackelberg payoff state-by-state. ■

**Example 4.** Suppose there is a single state, $|S_Q| = 1$. Suppose each player has three actions, $A_Q = \{T, M, B\}$, $A_{\text{SR}} = \{L, M, R\}$. Suppose – for ease of exposition – that payoffs are deterministic and as given in Figure 4. Assume the signalling structure $(\Phi, p)$ has full support. Suppose the short-run players do not observe past interactions so that a strategy for $\text{SR}^t$ is a map

$$\sigma_{\text{SR}^t} : \Phi \to \Delta(A_{\text{SR}}).$$

By an argument similar to the one in Case 2 of the proof of Theorem 4, the unique equilibrium payoff for the strategic long-run player is 0.

**Claim 2.** *There exist parameters for the algorithm $\langle Q^0, (\alpha^t), (\varepsilon^t)\rangle$ such that the $Q$-learning algorithm with these parameters achieves a payoff strictly above* 0.

*Proof.* Suppose $Q^0(T) = 2, Q^0(M) = 3$, and $Q^0(B) = 0$. Observe that there exists $x > 0$ such that $R$ is a best-response by the short-run player to a strategy $\tau \in \Delta(A_Q)$ only if $\tau(B) \geq x$. Assume that the experimentation rates $\varepsilon^t$ are low enough such that, given the signalling structure, the posterior probability that $B$ was played after any signal $\phi$ is less than $x$ for all strategies $\tau$ that put probability at most $\varepsilon_1$ on $B$.
For any $T$, if $\sigma_{\text{SR}^t}(\phi) \in \Delta(\{L, M\})$ for all $\phi \in \Phi$ and $t \leq T$, then $Q^t(R) \leq 0$ and $Q^t(T) \geq 2, Q^t(M) \geq 1$ almost surely. Given $Q^0$ and the condition on the experimentation probabilities, for $\sigma_{\text{SR}^t}$ puts probability 0 on $R$ after any signal $\phi$. Moreover, as long as

$$\mathbb{P}[Q^t(B) < \max\{Q^t(T), Q^t(M)\}] \leq x'$$

for some $x' > 0$ small enough, $\sigma_{\text{SR}^{t+1}}$ puts probability 0 on $R$ after any signal $\phi$. This, however, implies that $Q^t(B) < \max\{Q^t(T), Q^t(M)\}$ for all $t$ almost surely. As a consequence, $\sigma_{\text{SR}^{t+1}}$ puts probability 0 on $R$ after any signal $\phi$ for each period $t$.
Consequently, the expected payoff of the algorithm in period $t$ is weakly larger than $1 - 6\varepsilon^t$. The claim follows. □

We remark that the conclusions hold if we perturb the signalling structure or the payoffs. Moreover, payoffs can be random with their expectation given as in Figure 4. ■

The examples highlight illustrate the nuanced role commitment to an algorithm plays. In Example 3, the algorithm achieves a lower payoff than the strategic long-run player because the algorithm learns a myopic best-response. In Example 4, the long-run player achieves a lower payoff than the algorithm because, absent dynamic incentives and commitment power, the strategic long-run player plays a best-response to the short-run players' strategy in almost all periods. In equilibrium, the strategic player knows the action chosen by the short-run players and hence can choose a best-response to it, even if the short-run player's behavior changes from one period to the next. However, the algorithm need not play a best-response when the short-run player's behavior differs across periods.

## A.7. Auxiliary lemmas

### A.7.1.

Let $\{\mathcal{F}_\sqcup\}_t$ be a filtration on a probability space. For an index set $B \subset [0, 1]$, let $\left(v_t^b\right)_{t,b\in B}$ be a collection of independent random variables. Assume that $(v_s^b)_{b,s\geq k}$

are independent of $\mathcal{F}_t \forall k, t \leq k$. For each $t$, let $\chi_t^b$ be a collection of $\mathcal{F}_t$-measurable random variables with $\chi_t^b \in [0,1]$ and $\int_B \chi_t^b \, \mathrm{d}b = 1$ a.s.

Set $w_t = \int_{b \in B} \chi_t^b v_{t+1}^b \, \mathrm{d}b$. $w_t$ is $\mathcal{F}_{t+1}$-measurable. For ease of exposition, assume $B$ is finite so that $w_t = \sum_{b \in B} \chi_t^b \cdot v_{t+1}^b$.

Let $(\alpha^t)$ be a sequence in $(0,1)$ that satisfies Assumption (Step-Size). Define a stochastic process $\{W_t\}$ by

$$W_0 = 0 \quad \text{a.s.},$$
$$W_{t+1} = \left(1 - \alpha^t\right) W_t + \alpha^t w_t.$$

**Lemma 5.** *Assume* $\sup_{t,b \in B} \mathbb{E}[(v_t^b)^2] = \overline{\sigma^2} < \infty$. *For all* $\xi > 0$, *there exists a random time* $T_\xi$ *such that*

$$W_t \geq \inf_{t,b \in B} \mathbb{E}[v_t^b] - \xi$$

*for all* $t \geq T_\xi$, *and* $T_\xi < \infty$ *with probability* 1.

*Proof*Step 1 Assume $\mathbb{E}[v_t^b] = 0 \; \forall b, t$. Then

$$\mathbb{E}[w_t | \mathcal{F}_t] = \sum_{b \in B} \chi_t^b \mathbb{E}[v_{t+1}^b | \mathcal{F}_t] = 0.$$

The first equality holds because $\chi_t^b$ is $\mathcal{F}_t$–measurable and the second equality holds because $v_{t+1}^a$ is independent of $\mathcal{F}_t$ so that $\mathbb{E}[v_{t+1}^b | \mathcal{F}_t] = \mathbb{E}[v_{t+1}^b] = 0$. By a similar argument,

$$\mathbb{E}[w_t^2 | \mathcal{F}_t] = \sum_b \left(\chi_t^b\right)^2 \mathbb{E}[(v_{t+1}^b)^2] \leq \sup_b \mathbb{E}[(v_{t+1}^b)^2] \leq \overline{\sigma}^2.$$

Hence, $\lim_{t \to \infty} W_t = 0$ with probability 1 by, for example, Lemma 1, p. 190, in Tsitsiklis (1994).

Step 2 Consider the process

$$\tilde{W}_{t+1} = (1 - \alpha^t)\tilde{W}_t + \alpha^t \tilde{w}_t \quad \tilde{w}_t \equiv w_t - \mathbb{E}[w_t | \mathcal{F}_t].$$

Clearly, $\mathbb{E}[\tilde{w}_t | \mathcal{F}_t] = 0$ so that, by Step 1, $\tilde{W}_t \to 0$ with probability 1. Observe that, for every $t$,

$$\tilde{W}_{t+1} = W_{t+1} - \sum_{s \leq t} \beta_{s,t} \mathbb{E}[w_s | \mathcal{F}_s]$$

for deterministic scalars $\beta_{s,t} \geq 0, \sum_{s \leq t} \beta_{s,t} \leq 1$. Because $\sum_t \alpha^t = \infty$,

$\lim_{t\to\infty}\sum_{s\leq t}\beta_{s,t}=1$. Moreover, for each $t$,

$$\mathbb{E}[w_t|\mathcal{F}_t]=\sum_b \chi_t^b\mathbb{E}[v_{t+1}^b|\mathcal{F}_t]=\sum_b \chi_t^b\mathbb{E}[v_{t+1}^b]\geq \inf_a \mathbb{E}[v_{t+1}^b]$$

with probability 1. The second inequality holds because $v_{t+1}^b$ is independent of $\mathcal{F}_t$. By an analogous argument, $\mathbb{E}[w_t|\mathcal{F}_t]\leq \sup_b \mathbb{E}[v_{t+1}^b]$.
Fix $\xi>0$. Because $\tilde{W}_t\to 0$, there exists (a random variable) $\tilde{T}_{\xi/2}$ such that $\tilde{W}_t\geq -\xi/2$ for $t\geq \tilde{T}_{\xi/2}$. Let $\hat{T}$ be such that $\sum_{s\leq t}\beta_{s,t}\inf_{b\in B}\mathbb{E}[v_{s+1}^b]\geq \inf_{t,b\in B}\mathbb{E}[v_{t+1}^b]-\xi/2$ for $t\geq\hat{T}$. It follows that

$$W_t\geq \inf_{t,b}\mathbb{E}[v_t^b]-\xi$$

for $t\geq T_\xi\equiv\max\{\hat{T},\tilde{T}_{\xi/2}\}$. Moreover, $T_\xi<\infty$ with probability 1.

□

**Remark 2.** *The assumption that $B$ is finite can be relaxed.*

### A.7.2.

**Lemma 6.** *Suppose $(w_t)_t$ are independent, $\mathbb{E}[w_t]=0$, $w_t$ is sub-Gaussian for each $t$ with norm $||w_t||_{\psi_2}=K_t$.*[59] *Suppose $\sup_t K_t<\infty$. Assume $(\alpha^t)$ satisfies Assumption (Step-Size).*
*For a $W_0$ with $\mathbb{E}[W_0^2]<\infty$, define the process $(W_t)_t$ recursively by*

$$W_{t+1}=(1-\alpha^t)W_t+\alpha^t w_t.$$

*Fix $\delta>0$. Then, for every $\xi>0$ there exists $M\in\mathbb{N}$ such that if $W_M=0$ then*

$$\mathbb{P}[\{\textit{for some } t, |W_{M+t}|\geq\delta\}]<\xi.$$

*Proof.* First, note that if $W_M=0$, then $W_{M+t}=\sum_{i=1}^t\beta_i(M,t)w_{M+i}$ for some constants $\beta_i(M,t)>0, \sum_{i=1}^t\beta_i(M,t)\leq 1$. Note that the $\beta_i(M,t)$ depend on $M$ and $t$. By the assumption that $\alpha^{t+1}\geq\alpha^t(1-\alpha^{t+1})$, for fixed $M,t$, it is $\beta_i(M,t)\leq\alpha_{M+t}$.

Second, since the $(w_t)$ are independent, have mean 0 and are sub-Gaussian, we

[59] Recall that the sub-Gaussian norm is defined as

$$||w||_{\psi_2}=\inf\{c>0|\mathbb{E}[\exp(w^2/c^2)]\leq 2\}.$$

can apply Hoeffding's Inequality:[60] there exists a constant $c > 0$ such that,

$$\mathbb{P}[|W_{M+t}| \geq \delta] \leq 2 \exp\left(-\frac{c\delta^2}{\sum_{i=1}^{t} \beta_i^2(M,t)}\right).$$

Third, note that, for each $t, M$, $\sum_{i=1}^{t} \beta_i^2(M,t) \leq \frac{1}{\alpha_{M+t}} \alpha_{M+t}^2 = \alpha_{M+t}$.
Hence, for a constant $c(\gamma)$, independent of $M$,

$$\mathbb{P}[\{\text{for some } t, |W_{M+t}| \geq \gamma\}] \leq c(\gamma) \sum_{t=M}^{\infty} \exp(\left(-\frac{1}{\alpha^t}\right).$$

Note that for every $s > 0$ and $t$ large enough, $\exp(-1/\alpha^t) < \exp(-s\log(t))$. To see this, suppose that $\exp(-1/\alpha^t) \geq \exp(-s\log(t))$.

$$\begin{aligned}
&\exp(-1/\alpha^t) \geq \exp(-s\log(t)) \\
\iff &-\frac{1}{\alpha^t} \geq -s\log(t) \\
\iff &\frac{1}{s^2\log^2(t)} \leq \alpha^{t2} \\
\implies &\sum_t \alpha^{t2} \geq \sum_t \frac{1}{s^2\log^2(t)} \geq \sum_t \frac{1}{s^2 t} = \infty,
\end{aligned}$$

a contradiction to Assumption (Step-Size).
Consequently,

$$\mathbb{P}[\{\text{for some } t, |W_{M+t}| \geq \gamma\}] \leq c(\gamma) \sum_{t=M}^{\infty} \exp\left(-2\log(t)\right) = c(\gamma) \sum_{t=M}^{\infty} \frac{1}{t^2} < \infty.$$

□

# References


Abada, Ibrahim and Xavier Lambin (2023). "Artificial Intelligence: Can seemingly collusive outcomes be avoided?" In: *Management Science* 69.9, pp. 5042–5065.

Arslan, Gürdal and Serdar Yüksel (2016). "Decentralized Q-learning for stochastic teams and games". In: *IEEE Transactions on Automatic Control* 62.4, pp. 1545–1558.


[60] See Theorem 2.6.3, p. 27, in Vershynin (2018).

Asker, John, Chaim Fershtman, and Ariel Pakes (2022). "Artificial Intelligence, Algorithm Design, and Pricing". In: *AEA Papers and Proceedings*. Vol. 112, pp. 452–56.

— (2023). "The Impact of Artificial Intelligence Design on Pricing". In: *Journal of Economics & Management Strategy*.

Bagwell, Kyle (1995). "Commitment and Observability in Games". In: *Games and Economic Behavior* 8.2, pp. 271–280.

Balcan, Maria-Florina et al. (2015). "Commitment without regrets: Online learning in stackelberg security games". In: *Proceedings of the sixteenth ACM conference on economics and computation*, pp. 61–78.

Banchio, Martino and Giacomo Mantegazza (2023). "Adaptive Algorithms and Spontaneous Collusion". In: *arXiv preprint arXiv:2202.05946*.

Barberis, Nicholas and Lawrence Jin (2023). *Model-free and model-based learning as joint drivers of investor behavior*. Working Paper 31081. National Bureau of Economic Research.

Bertrand, Quentin et al. (2023). "Q-learners Can Provably Collude in the Iterated Prisoner's Dilemma".

Bhaskar, Venkataraman and Eric van Damme (2002). "Moral Hazard and Private Monitoring". In: *Journal of Economic Theory* 102.1, pp. 16–39.

Blum, Avrim, Nika Haghtalab, and Ariel Procaccia (2014). "Learning optimal commitment to overcome insecurity". In: *Advances in Neural Information Processing Systems* 27.

Boer, Arnoud den, Janusz Meylahn, and Maarten Pieter Schinkel (2022). "Artificial collusion: Examining supracompetitive pricing by Q-learning algorithms". In: *Amsterdam Law School Research Paper* 2022-25.

Börgers, Tilman and Rajiv Sarin (1997). "Learning through reinforcement and replicator dynamics". In: *Journal of Economic Theory* 77.1.

Braverman, Mark et al. (2018). "Selling to a no-regret buyer". In: *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 523–538.

Brown, Zach and Alexander MacKay (2023a). "Collusion and Coercion with Naive Rivals".

— (2023b). "Competition in Pricing Algorithms". In: *American Economic Journal: Microeconomics* 15.2, pp. 109–156.

Brückner, Michael and Tobias Scheffer (2011). "Stackelberg games for adversarial prediction problems". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 547–555.

Calvano, Emilio et al. (2019). "Algorithmic pricing what implications for competition policy?" In: *Review of Industrial Organization* 55, pp. 155–171.

— (2020). "Artificial Intelligence, Algorithmic Pricing, and Collusion". In: *American Economic Review* 110.10, pp. 3267–3297.

Calvano, Emilio et al. (2021). "Algorithmic collusion with imperfect monitoring". In: *International Journal of Industrial Organization* 79, p. 102712.

Camara, Modibo, Jason Hartline, and Aleck Johnsen (2020). "Mechanisms for a no-regret agent: Beyond the common prior". In: *2020 IEEE 61st Annual Symposium on Foundations of Computer Science*. IEEE, pp. 259–270.

Cartea, Alvaro et al. (2022). "The Algorithmic Learning Equations: Evolving Strategies in Dynamic Games". In: *available at SSRN 4175239*.

D'Andrea, Maurizio (2023). "Playing against no-regret players". In: *Operations Research Letters* 51.2, pp. 142–145.

Decarolis, Francesco et al. (2023). "Artificial Intelligence & Data Obfuscation: Algorithmic Competition in Digital Ad Auctions". mimeo.

Deng, Yuan, Jon Schneider, and Balasubramanian Sivan (2019). "Strategizing against no-regret learners". In: *Advances in neural information processing systems* 32.

Dolgopolov, Arthur (2024). "Reinforcement Learning in a Prisoner's Dilemma". In: *Games and Economic Behavior* 144, pp. 84–103.

Dorner, Florian (2021). "Algorithmic collusion: A critical review".

Fiez, Tanner, Benjamin Chasnov, and Lillian Ratliff (2019). "Convergence of learning dynamics in stackelberg games". In: *arXiv preprint arXiv:1906.01217*.

Fudenberg, Drew and David Levine (1998). *The Theory of Learning in Games*. Vol. 2. MIT press.

Guruganesh, Guru et al. (2024). "Contracting with a learning agent". In: *arXiv preprint arXiv:2401.16198*.

Haghtalab, Nika, Chara Podimata, and Kunhe Yang (2023). "Calibrated Stackelberg Games: Learning Optimal Commitments Against Calibrated Agents". In: *arXiv preprint arXiv:2306.02704*.

Hansen, Karsten, Kanishka Misra, and Mallesh Pai (2021). "Algorithmic collusion: Supra-competitive prices via independent algorithms". In: *Marketing Science* 40.1, pp. 1–12.

Hart, Sergiu (1992). "Games in Extensive and Strategic Forms". In: *Handbook of Game Theory with Economic Applications*. Ed. by Robert J. Aumann and Sergiu Hart. Vol. 1. Amsterdam: North Holland: Elsevier. Chap. 2, pp. 19–40.

Hart, Sergiu and Andreu Mas-Colell (2013). *Simple Adaptive Strategies: From Regret-Matching to Uncoupled Dynamics*. World Scientific Publishing.

Hettich, Matthias (2021). "Algorithmic Collusion: Insights from Deep Learning".

Johnson, Justin, Andrew Rhodes, and Matthijs Wildenbeest (2023). "Platform Design when Sellers use Pricing Algorithms". In: *Econometrica* 91.5, pp. 1841–1879.

— (2024). "Algorithmic Steering and Advertising on Platforms".

Kalai, Ehud and Ehud Lehrer (1993). "Rational learning leads to Nash equilibrium". In: *Econometrica*, pp. 1019–1045.

Kephart, Jeffrey, James Hanson, and Amy Greenwald (2000). "Dynamic pricing by software agents". In: *Computer Networks* 32.6, pp. 731–752.

Kianercy, Ardeshir and Aram Galstyan (2012). "Dynamics of Boltzmann $Q$-learning in two-player two-action games". In: *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* 85.4, p. 041145.

Klein, Timo (2021). "Autonomous algorithmic collusion: Q-learning under sequential pricing". In: *The RAND Journal of Economics* 52.3, pp. 538–558.

Lamba, Rohit and Sergey Zhuk (2022). "Pricing with Algorithms". In: *arXiv preprint arXiv:2205.04661*.

— (in progress). "Algorithmic Pricing $\implies$ Supracompetitive Pricing?"

Lambin, Xavier (2024). "Less than meets the eye: simultaneous experiments as a source of algorithmic seeming collusion". In: *available at SSRN 4498926*.

Leslie, David and Edmund Collins (2005). "Individual Q-learning in normal form games". In: *SIAM Journal on Control and Optimization* 44.2, pp. 495–514.

Letchford, Joshua, Vincent Conitzer, and Kamesh Munagala (2009). "Learning and approximating the optimal strategy to commit to". In: *Algorithmic Game Theory: Second International Symposium, SAGT 2009, Paphos, Cyprus, October 18-20, 2009. Proceedings 2*. Springer, pp. 250–262.

Levine, David (2023). "Efficiently Breaking the Folk Theorem by Reliably Communicating Long Term Commitments".

Maggi, Giovanni (1999). "The value of commitment with imperfect observability and private information". In: *The RAND Journal of Economics* 30.4, pp. 555–574.

Mailath, George and Larry Samuelson (2014). "Reputations in Repeated Games". In: *Handbook of Game Theory with Economic Applications*. Ed. by Peyton Young and Shmuel Zamir. Vol. 4. Amsterdam: North Holland: Elsevier. Chap. 4, pp. 165–238.

Mansour, Yishay et al. (2022). "Strategizing against learners in Bayesian games". In: *Conference on Learning Theory*. PMLR, pp. 5221–5252.

Marecki, Janusz, Gerry Tesauro, and Richard Segal (2012). "Playing repeated stackelberg games with unknown opponents". In: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pp. 821–828.

Matsushima, Hitoshi (1991). "On the theory of repeated games with private information: Part I: anti-folk theorem without communication". In: *Economics Letters* 35.3, pp. 253–256.

Nachbar, John (1997). "Prediction, optimization, and learning in repeated games". In: *Econometrica*, pp. 275–309.

Possnig, Clemens (2024). "Reinforcement Learning and Collusion".
Qiu, Liying et al. (2023). "Does Personalization in Product Rankings Facilitate or Mitigate Algorithmic Pricing Collusion?" In: *available at SSRN 4132555*.
Robbins, Herbert and Sutton Monro (1951). "A stochastic approximation method". In: *The Annals of Mathematical Statistics* 22.3, pp. 400–407.
Rodrigues Gomes, Eduardo and Ryszard Kowalczyk (2009). "Dynamic analysis of multiagent Q-learning with $\varepsilon$-greedy exploration". In: *Proceedings of the 26th annual international conference on machine learning*, pp. 369–376.
Salcedo, Bruno (2015). "Pricing Algorithms and Tacit Collusion". mimeo.
Sandholm, Tuomas and Robert Crites (1996). "Multiagent reinforcement learning in the iterated prisoner's dilemma". In: *Biosystems* 37.1-2, pp. 147–166.
Schäfer, Maximilian (2022). "On the Emergence of Cooperation in the Repeated Prisoner's Dilemma". In: *arXiv preprint arXiv:2211.15331*.
Sutton, Richard and Andrew Barto (2018). *Reinforcement Learning: An Introduction*. 2nd ed. Cambridge, MA: MIT press.
Tesauro, Gerald and Jeffrey Kephart (2002). "Pricing in agent economies using multi-agent Q-learning". In: *Autonomous agents and multi-agent systems* 5, pp. 289–304.
Tsitsiklis, John (1994). "Asynchronous Stochastic Approximation and Q-learning". In: *Machine learning* 16, pp. 185–202.
Van Damme, Eric and Sjaak Hurkens (1997). "Games with Imperfectly Observable Commitment". In: *Games and Economic Behavior* 21.1-2, pp. 282–308.
Vershynin, Roman (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press.
Waizmann, Stephan (2024). "Agent vs Algorithm: Repeated Play against a $Q$-Learning Algorithm". in progress.
Waltman, Ludo and Uzay Kaymak (2008). "Q-learning agents in a Cournot oligopoly model". In: *Journal of Economic Dynamics and Control* 32.10, pp. 3275–3293.
Wang, Qiaochu et al. (2023). "Algorithms, artificial intelligence and simple rule based pricing". In: *available at SSRN 4144905*.
Watkins, Christopher (1989). "Learning from Delayed Rewards". Ph.D. Dissertation. Cambridge, United Kingdom: King's College, Cambridge University.
Watkins, Christopher and Peter Dayan (1992). "Q-learning". In: *Machine Learning* 8, pp. 279–292.
Werner, Tobias (2023). *Algorithmic and human collusion*. mimeo.
Wiseman, Thomas (2005). "A Partial Folk Theorem for Games with Unknown Payoff Distributions". In: *Econometrica* 73.2, pp. 629–645.
Yang, Guosong, Radha Poovendran, and Joao Hespanha (2019). "Adaptive learning in two-player Stackelberg games with continuous action sets". In: *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, pp. 6905–6911.

Young, Peyton (2004). *Strategic Learning and its Limits*. Oxford, United Kingdom: Oxford University Press.

Zhao, Geng et al. (2023). "Online learning in stackelberg games with an omniscient follower". In: *International Conference on Machine Learning*. PMLR, pp. 42304–42316.

Zrnic, Tijana et al. (2021). "Who leads and who follows in strategic classification?" In: *Advances in Neural Information Processing Systems* 34, pp. 15257–15269.