Exploration, Exploitation, Amelioration: Experimentation with Endogenously Changing Arms

Evgeniya Kudinova*

November 2023

Click here for the latest version

Abstract

I analyse how economic agents adapt to new risky opportunities, such as new technologies, when the agents can invest in increasing the likelihood of a successful outcome, while also learn about its original quality. I build on a single risky arm Poisson bandit environment and explore how the ability to endogenously change the arm, by investing, affects the incentives for experimentation. More specifically, I assume that successful investment turns a bad arm into a good one.

As opposed to standard good news Poisson bandits, I find that beliefs may evolve non-monotonically and that the agent may get stuck on a certain belief and invest at some intensity until the news arrives. In the context of adaptation, this means that the agent may keep pursuing the new opportunity forever despite not reaching a success for a long time, in contrast to necessarily giving up according to the traditional experimentation models. This creates discontinuity in the long-term outcomes of experimentation and suggests strong implications for innovation design and organisational strategies for technological adaptation.

*Department of Economics, London School of Economics. E-mail: e.kudinova@lse.ac.uk. I am deeply grateful to Gilat Levy for her invaluable guidance, support and encouragement throughout this project completion. I would also like to extend special gratitude to Ronny Razin and Stephane Wolton for insightful discussions, constructive suggestions and their general support. For very helpful comments I also thank Thomas Brzustowski, Matthias Doepke, Andrew Ellis, Maitreesh Ghatak, Ethan Ilzetzki, Alastair Langtry, Hongting Leng, Matt Levy, Alan Manning, Ben Moll, Ines Moreno de Barreda, Dmitry Mukhin, Francesco Nava, Martin Pesendorfer, Dimitra Petropoulou, Christopher Sandmann, Marcia Schafgans, Balazs Szentes, John Van Reenen, Cecilia Wood, Leeat Yariv, Alwyn Young, Alfred Zhang and numerous seminar participants at LSE. Any mistakes remain my own.

1 Introduction

Experimentation, e.g., with a new technology, is an inevitable step towards achieving progress. Yet, it involves a fundamental risk as the new technology may potentially turn out to be worse than the current one. Despite this opportunity cost, the possibility to learn that the new technology is indeed beneficial serves as a strong incentive to start and prolong experimentation at least for a while.

Due to the importance of experimentation for economic progress and the strategic nature of the problem, a large literature has analysed the optimal decision of experimenters. One aspect that the traditional models typically disregard, however, is the potential ability of decision makers to affect the outcome of experimentation instead of just learn about it. Indeed, in many applications the agents may actively engage in adapting to the new risky opportunities, and thus troubleshoot any issues on the way to ensure that the outcome is a success. A firm adopting a new technology may assign a team to monitor how it rolls out in the specific company environment and tailor it to become a better fit. Governments launching a new policy typically employ bureaucrats to achieve successful implementation. A worker, switching to a new occupation and realizing that she may lack some necessary skills, may attend advanced training courses, while a team leader hiring a new trainee of uncertain quality can invest in training and supervising the apprentice. These examples raise the question of when investing in an adaptation process - an investment that is potentially costly in terms of time, effort or physical resources - is worthy to the agent. In this paper I show that being able to make such investments has a meaningful impact on experimentation and learning, stemming beyond the direct benefits thereof.

I analyse experimentation with the ability to actively change the risky option (I will refer to this process as training). Specifically, I augment the canonical Poisson bandit model, as formulated by Keller et al. (2005) (henceforth KRC). There is a single decision maker (or agent) who at each instance allocates her time between exploiting a safe arm that yields a fixed known payoff and experimenting using a risky one. The risky arm can be good or bad,

where the good one achieves an observable breakthrough (also referred to as good news) at a Poisson arrival rate and is preferred to the safe one, while the bad one can never provide any positive payoff. On top of this traditional setup, I assume that, conditionally on using it, the agent can choose to endogenously change the risky arm: By paying some cost, she can turn a bad risky arm into a good one at some Poisson arrival rate; that is, if the agent invests in training, there is some probability that the bad risky arm becomes good, while a good arm remains unaffected. I further assume that once the change has occurred it is irreversible. Importantly, the event that training was successful is not directly observable to the agent; she can learn only through observing the breakthrough. [1]

The ability to invest in training the risky arm leads to significant qualitative changes compared to the case in which this is not possible. Specifically, training affects the evolution of beliefs about the probability that the risky arm is good. While the traditional Poisson bandits with good news imply that the agent necessarily gets more pessimistic in the absence of news arrival reading it as a signal of a bad arm, here the beliefs may evolve in a non-monotone manner. In fact, the positive effect of investing in training may counteract and even exceed the traditional negative effect on beliefs from receiving no news. I show that if the agent invests in improvements, she may become more optimistic even in the absence of news under relatively low priors, or still more pessimistic under high ones. Such non-monotonicity in beliefs has a large impact on the agent's optimal strategies and the outcomes she can achieve.

My key result, derived from the non-monotonicity of belief evolution described above, is that the absence of good news does not necessarily lead to quitting experimentation. This is a sharp contrast to the standard models where the agent surely gives up on experimenting upon getting no news for a while, as then her beliefs gradually drop until they reach a stopping

[1] I am considering the ability to improve the arm only conditional on using it. Such assumption fits the motivating examples and the story of investments in adaptation, when the agents learn more once they start the risky endeavor and can troubleshoot on the way. This is in contrast to making investments before the experimentation starts, which would be reflected through adjusting the prior belief on the arm being good, but become strictly dominated once experimentation starts, hence not affecting the analysis of this paper.

cutoff. In turn, when the ability to train guarantees a sufficiently high likelihood to improve the arm, once the agent starts training the arm, she will keep doing so 'forever' at least at some intensity even if no news arrives. This arises when the positive effect of improving the risky arm dominates the negative one from getting no news around the stopping cutoff (below which experimenting is no longer worthwhile), making the agent more optimistic irrespective of the informational flow. As such, the agent's belief diverges away from the cutoff deeper into the experimentation region, and the agent never quits experimenting: Her belief converges to the one where the two effects balance each other out, and the agent gets stuck training the arm until she finally observes the good news and resolves the uncertainty.

I also show that the overall optimal strategies structure may be non-monotone in actions. I find that the agent prefers to train the arm for some moderate beliefs, and may sometimes find it optimal to purely exploit the risky arm being both more optimistic and more pessimistic than in the training range. For lower productivity of training this leads to non-monotone actions path. Conditional on observing no news, the agent may first (for high enough priors) purely experiment for a bit, then train the risky arm as her beliefs go down and then again just experiment before switching to the safe option. This occurs because, as she anticipates quitting the risky arm soon, she no longer values its improvement enough to be willing to incur any extra cost.

For more efficient training, the non-monotone actions structure creates a sharp discontinuity in the agent's behaviour and the outcome of experimentation. Specifically, there exists a cutoff belief such that, when being slightly more pessimistic, the agent may give the risky arm a chance before exiting experimentation, but the path will necessarily lead her to switching to a safe option conditional on getting no news. In contrast, starting slightly above such belief, she will only become more optimistic and eventually get stuck on training the arm at certain intensity until news arrive. This ensures that she obtains a breakthrough almost surely in the limit, as opposed to much less optimistic outcome in the former scenario.

Overall, my results highlight that the ability to train impacts experimentation beyond

its direct effect on the bad arms. In fact, training extends the experimenting horizon for the agent, which increases the likelihood of discovering initially good arms and reduces the chance of erroneous abandonment of those - an event occurring with strictly positive probability in traditional experimentation frameworks. The strength of these effects varies in a discontinuous manner given the variation in the agent's optimal actions paths.

My findings suggest important implications for the prior design of the arms and training mechanisms to anyone pursuing the goal of spreading risky arms adoption, or innovation more generally. I show that a small change in the productivity of training can cause a discontinuous switch from the pessimistic regime, where risky arms may be abandoned after a while, to the optimistic one, with all risky arms resulting in a breakthrough. Similarly, a marginal improvement of the original technology (arm) may lead to discontinuously higher investments in training and a better spread of innovation in the long run.

The change in action dynamics due to training also leads to novel technical implications. Firstly, the resulting value function does not have to be smooth in contrast to the standard good news experimentation models. This is due to the existence of a cutoff that the agent diverges away from being both above and below it. The threshold belief is then defined purely by *continuous pasting* as opposed to the smooth pasting argument, following the terminology introduced by Keller and Rady (2015). [2] Secondly, the cutoff belief where the agent gets stuck being indifferent between experimenting with or without training is absorbing on both sides: A more pessimistic agent will train and increase her belief until she converges to the cutoff, while a more optimistic one will exploit the risky arm and become more pessimistic. Such fully absorbing cutoffs are novel in the experimentation literature, and I establish an approach for defining these types of boundaries, which requires an extra optimality condition on top of traditional value matching and smooth pasting ones.

The rest of the paper is organized as follows. The next section discusses related literature

[2]Their paper documents a similar diverging behaviour at the stopping cutoff in the experimentation with breakdowns, where the beliefs move away from the cutoff on one side and remain constant forever on the other. There, the benefit of learning the arm's type around the cutoff is null given that it leads to switching to the safe arm, an intuition that is very different from the one I identify in this paper.

and emphasizes the contribution of this paper. In Section 3 I present the model, analyse the evolution of beliefs and the value function, and provide the benchmark solution when the agent is not allowed to train the risky arm. I then establish the main results and qualitative implications of those in Section 4. In Section 5 I discuss the impact of the experimentation outcomes divergence in the long run and extend the baseline model to 1) a more general information structure when the agent can observe both good and bad news and 2) an environment where the agent can vary the training efficiency at some convex cost. Section 6 concludes by providing a broader perspective on how the ability to train changes the learning from experimenting.

2 Related literature

This paper contributes to the rich literature on the experimentation with Poisson bandits, pioneered by KRC, and explored further in Keller and Rady (2010, 2015), Klein and Rady (2011), Guo (2016), Yariv (2021) and others. In all above, the risky arms types are fixed and stable. There are papers that analyse *restless* bandits, where the arms are allowed to randomly change in time irrespective of whether being used, as in Whittle (1988) or Keller and Rady (1999). Safronov (2021) considers agents who can learn by doing, that is, increase the lump sum payoff of the risky arms via becoming experienced in exploiting a certain option. Fryer and Harms (2018) analyse the environment where the risky arm's payoff increases if in use and decreases otherwise. In both of these papers, as in the classic restless arms models, the arms change exogenously once the agent decides to use them. In contrast, in the environment I study in this paper the agent actively chooses whether she wants the arm to change and this decision is separate from the exploitation one.

Quite a few papers consider experimentation subject to moral hazard. They typically focus on the principal-agent environments where the principal is providing incentives for the agent to exert effort for conducting experimentation. In such or similar setups, Bergemann

and Hege (1998, 2005), Hörner and Samuelson (2013) and Hidir (2019) study various aspects of optimal contracting to incentivize experimentation by a single agent. Halac et al. (2016) introduce adverse selection on top of the moral hazard to the contracting problem. Halac et al. (2017) and Moroni (2022) extend the optimal incentives provision design further to contest-like environments where multiple agents compete in experimentation over the same tasks. Diverting from contractual design, Bonatti and Hörner (2011) analyse a model of team experimentation with binary type risky arm, where the arrival of breakthrough depends on the joint effort of the experimenting agents. In all these papers, if considered with a single agent, effort exertion transfers a safe arm into a risky one (determines the type of a good risky arm), or, in other words, gives rise to experimentation. This differs from the model discussed here, where improving the arm with training comes on top of experimentation, and the experimentation problem preserves for any training decision.

Finally, Fershtman and Pavan (2023) consider a decision maker who chooses between experimenting on some known arms and searching for the new ones at each instance. That is, they endogenize the set of arms available to the agent. An example of their framework can include discovering a new (and hence independent) exogenously changing arm in addition to the already available safe and fixed risky ones, and choosing one of the, now three, alternatives. Endogenizing the risky arm, as I propose in this paper, is very different from the three arms interpretation outlined above and thus is not nested there. It would require perfect correlation between the fixed and the restless risky arms, so any learning or pulling the changing arm affects them equally in both the belief and the underlying type.

3 Model

3.1 Baseline environment

Consider a single decision maker (agent) who, at each moment of time, chooses whether to use a safe or a risky arm. The risky arm is of a binary type $\lambda^\theta \in \{0, \lambda\}$, with $\lambda > 0$. If used

in $[t, t+dt)$, such arm generates a payoff normalized to 1 with probability $\lambda^\theta dt$ conditional on its type. That is, the payoffs (breakthroughs) are generated at a Poisson arrival rate λ^θ. The arm's type is unobserved by the decision maker, and she has a prior that the arm is 'good' with some probability p_0, i.e. $\Pr(\lambda^\theta = \lambda) = p_0$. Alternatively, the agent can use a safe arm, which generates a constant payoff $s \in (0, \lambda)$, so that the good risky arm is preferred to the safe one, while the bad one is dominated.

Denote the share of time/effort allocated to risky arm in $[t, dt)$ as $\alpha_t \in [0, 1]$ and the posterior belief about the risky arm's type conditional on the information available at time t as p_t, i.e. $\Pr(\lambda^\theta = \lambda | I_t) = p_t$. Then, the expected payoff generated in $[t, dt)$ is equal to $\big((1-\alpha_t)s + \alpha_t p_t \lambda\big)dt$, and the agent chooses $\alpha_t \in [0,1]_0^\infty$ to maximize the discounted present value of payoffs flow:

$$\max_{(\alpha_t \in [0,1])_0^\infty} E_{p_0}\left[\int_0^\infty e^{-rt}\big((1-\alpha_t)s + \alpha_t p_t \lambda\big)dt\right],$$

where $r > 0$ is a rate of exponential discounting.

3.2 Training/investing mechanism

I now add an additional ingredient to the model: suppose that the decision maker can also invest in improving the risky arm (that is, 'train') while exploiting it. Specifically, I assume that the training mechanism allows to turn a bad type of the arm into a good one (while it has no effect on the initially good arm). The bad arm becomes good at some Poisson arrival rate $\pi > 0$, i.e. with probability πdt if trained in $[t, t+dt)$. I refer to the event of the arm's type switch as 'successful training', and assume that this event is irreversible (that is, once the bad arm turns good, it remains good forever) and not directly observable (i.e., the agent cannot observe when the training succeeded directly, but only when the payoffs/breakthroughs are generated). The training mechanism is costly with a fixed cost $\kappa \geq 0$, which is payable no matter whether the training is successful or not; hence, if training

occurs in $[t, t+dt)$, the cost κdt is paid.

I still denote by $\alpha_t \in [0,1]$ the intensity of experimentation in $[t, t+dt)$ independently of whether training arises or not. However, the agent now also decides whether she wants to train the arm, and I introduce this decision through the intensity of training in $[t, t+dt)$, $\beta_t \in [0,1]$. [3] Then, the decision maker's maximization problem becomes:

$$\max_{((\alpha_t,\beta_t)\in[0,1]^2)_0^\infty} E_{p_0}\left[\int_0^\infty e^{-rt}\left((1-\alpha_t)s + \alpha_t\left(p_t\lambda - \beta_t\kappa\right)\right)dt\right].$$

Note that the safe arm is used $1-\alpha_t$ of the time, the risky arm (without training) $\alpha_t(1-\beta_t)$, and the risky arm is trained and used $\alpha_t\beta_t$ share of the time. In addition, observe that the only direct effect of training on the payoffs is through its cost, while the benefit appears in the objective function only indirectly: It is built into the likelihood of the risky arm being good at time t, that is, p_t. [4]

3.3 Beliefs evolution

The way beliefs evolve depends on the intensities of experimentation and training, α_t, β_t. Under the exploitation of the safe arm, that is if $\alpha_t = 0$ in $[t, dt)$, there is no additional informational flow concerning the risky arm's type, meaning that beliefs remain unchanged, and so $p_{t+dt} = p_t$.

If experimentation occurs ($\alpha_t > 0$), the payoff flow from the risky arm signals the arm's type. Given the assumption of conclusive breakthroughs ($\lambda^\theta = 0$ for bad type), once a positive payoff is observed, the risky arm is certain to be good, that is belief discretely jumps upwards to $p_{t+dt} = 1$. No news instead serves as a negative 'signal' and leads to

[3] One way to think about β_t is as intensity of training in $[0,1]$. Alternatively, one can imagine $\beta_t \in \{0,1\}$. Such restricting assumption does not affect any of the results, just slightly changes the interpretation of some of them.

[4] Think about λ^θ being λ_t^θ instead. Indeed, with training, the risky arm's type is allowed to evolve endogenously rather than remain at λ_0^θ forever under regular experimentation as in KRC. As such, under training, posterior belief $p_t \equiv \Pr(\lambda_t^\theta = \lambda)$ evolves because of the possible λ_t^θ evolution on top of the regular learning effect.

gradual belief updating downwards. In particular, following the classical Poisson bandits literature, $p_{t+dt} - p_t \equiv dp_t = -\alpha_t \lambda p_t(1-p_t)dt$. The belief is updated downwards more if the difference between the risky arm's types is greater ($\Delta\lambda^\theta = \lambda$ here), the intensity of experimentation is larger, and the more uncertain the agent is (for moderate beliefs p_t).

Now, on top of the standard learning effect, there is potentially another effect on belief evolution due to training. If the arm is good, training is redundant; if the arm is bad instead (with probability $1-p_t$), there is a chance πdt that the training is successful and the arm type shifts to a good one. This creates a positive boost in the belief evolution of size $\beta_t \pi(1-p_t)dt$, which counteracts the negative effect of getting no news. Note that this positive effect is stronger the stronger is the intensity of training.

The following lemma summarizes how beliefs change:

Lemma 1 *For any p_t and $\alpha_t > 0$, the risky arm is believed to be certainly good conditional on positive payoff of size 1 generated by it: $p^{news}_{t+dt} = 1$. Conditional on no payoff generated by the risky arm, beliefs follow: $dp_t = \alpha_t(\beta_t \pi - \lambda p_t)(1-p_t)dt$.*

Notice that the ability to endogenously change the arm may result in non-monotone paths of p_t even conditional on no news observed, and the direction of belief evolution depends on the intensity of training β_t as well as the current belief p_t. Trivially, if the agent purely experiments with the risky arm ($\beta_t = 0$), having no news results in a gradual downward shift in beliefs, for any prior $p_0 \in (0,1)$ (as shown by dashed paths in the Figure 1). With training at some positive intensity (solid lines), the dynamics of belief varies depending on the prior optimism of an agent. If her belief p_0 is high ($p_0 > \beta\frac{\pi}{\lambda}$), she still becomes more pessimistic without the news, but training slows down the beliefs depreciation due to the boost of optimism it gives (plotted in blue), and the beliefs converge to $\beta\frac{\pi}{\lambda}$. In contrast, for relatively low beliefs $p_0 < \beta\frac{\pi}{\lambda}$ (red scenario), the informational effect of no news is weak enough that it is dominated by the positive influence of training. In fact, with continuous training the agent gets more optimistic, which gradually increases the informational content of getting no news and hence the strength of the negative effect on belief evolution up until

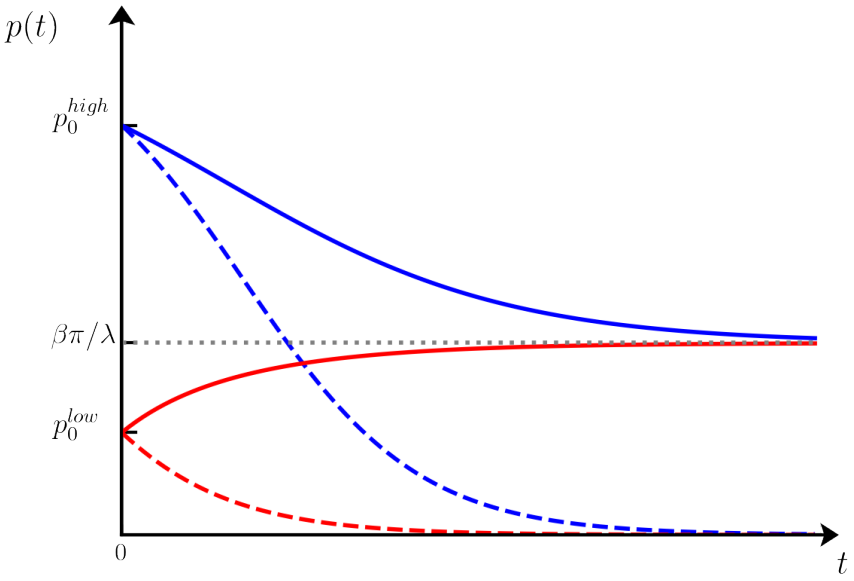

Figure 1: Evolution of beliefs conditional on receiving no news: for experimenting forever ($\alpha_t = 1, \beta_t = 0$) in dashed lines, and training forever at intensity $\beta_t = \beta > 0$ in solid ones; for high p_0 in blue, and low p_0 in red.

the point where the two effects counteract each other exactly. At such belief ($p_t = \beta_t \frac{\pi}{\lambda}$) the agent gets 'stuck' for a while, and the question arises whether and when such a path can be optimal.

3.4 Bellman equation and the value function

In this subsection I derive the Hamilton-Jacobi-Bellman equation (HJB) that characterises the maximisation problem of interest, as well as the value function obtained from solving the equation. Note that all the extra dynamics occurring due to the ability to endogenously change the arm is represented through the belief about the arm's type, that is p_t. This means that a single state variable captures the entire dynamics of the problem, which allows for a tractable closed-form solution.

Denote by $V(p_t)$ the value function resulting from the optimal control problem. It consist of the monetary payoff obtained in $[t, t + dt)$ and the discounted continuation value of the problem. With the probability of observing good news from the risky arm, belief jumps to

$p_{t+dt} = 1$ and otherwise changes according to the law of motion identified and discussed in section 3.3. Overall, it can be represented as:

$$V(p_t) = \max_{(\alpha_t,\beta_t)\in[0,1]^2} \big((1-\alpha_t)s + \alpha_t\left(p_t\lambda - \beta_t\kappa\right)\big)dt + e^{-rdt}\big(\alpha_t p_t\lambda dt V(1) + (1-\alpha_t p_t\lambda dt)V(p_t + dp_t)\big)$$

$$\text{s.t. } dp_t = \alpha_t(\beta_t\pi - \lambda p_t)(1-p_t)dt$$

Applying a first-order approximation on all the non-linear components, setting $(dt)^2$ terms to 0 and dividing the equation by dt (as well as dropping the time subscripts, since the time dimension is fully reflected through p_t) simplifies the problem further to:

$$rV(p) = \max_{(\alpha,\beta)\in[0,1]^2} \big((1-\alpha)s + \alpha\left(p\lambda - \beta\kappa\right)\big) + \alpha p\lambda\big(V(1) - V(p)\big) + V'(p)\alpha(\beta\pi - \lambda p)(1-p)$$

The equation shares a first-order differential-difference form as obtained in traditional Poisson bandit's models, and if the training option is shut down by setting $\beta = 0$, the equation is reduced to the KRC one. Even with $\beta > 0$, the interpretation of the HJB equation remains similar. The first part represents the current payoff obtained from picking a certain strategy. The latter parts describe the change in the value function due to the change in belief. If the good news is observed (at the rate $\alpha p\lambda$), there is a discrete jump in belief and value. Otherwise, at a rate approximated by 1, the belief shifts just marginally and incurs a marginal shift in the continuation value, $V'(p)$.

Yet, the ability to change the arm through β substantially changes the potential scope of the solutions as well as the resulting dynamics. Typically, the effects of acquiring new information move in opposite directions where the discrete jump from good news is positive, and the smooth shift in the absence of it is negative. In contrast, here the direction of the marginal shift is non-monotone and depends both on the state and the control variables value. That is, the agent endogenizes her learning from receiving no news, and it is possible that the continuation value increases with any information she gets (i.e. both informational

effects in the equation are positive).

Importantly, the HJB equation is linear in both control variables, α and β. This implies that it is optimal for the agent to take pure actions: use a safe arm, purely experiment, or experiment with training the arm at full intensity. As such, the equation can be solved separately for each of the three modes resulting in the following value functions:

$$V(p) = \max_{\alpha,\beta} \begin{cases} \frac{s}{r} & \alpha = 0 \text{ (safe arm)} \\ \frac{\lambda}{r}p + C_{Exp}f(p) & \alpha = 1, \beta = 0 \text{ (experimentation)} \\ y(p) + C_{Tr}g(p) & \alpha = \beta = 1 \text{ (training)} \end{cases} \tag{1}$$

where $y(p) = p\left(\frac{\lambda}{r} - \frac{\kappa}{r+\lambda}\right) + (1-p)\left(\frac{\pi}{r+\pi}\left(\frac{\lambda}{r} - \frac{\kappa}{r+\lambda}\right) - \frac{\kappa}{r+\pi}\right)$, $g(p) = \frac{(1-p)^{1+\frac{r+\pi}{\lambda-\pi}}}{|p-\frac{\pi}{\lambda}|^{\frac{r+\pi}{\lambda-\pi}}}$, $f(p) = \frac{(1-p)^{1+\frac{r}{\lambda}}}{p^{\frac{r}{\lambda}}}$, and $\{C_{Exp}, C_{Tr}\} \geq 0$ are some arbitrary constants.

The value from pure experimentation ($\alpha = 1, \beta = 0$) and from using the safe arm ($\alpha = 0$) exactly replicate the value functions from KRC. The agent gets $\frac{s}{r}$ forever under using the safe arm (and never switches to any other action since there is no dynamic belief updating), and $\frac{\lambda}{r}p$ if she experiments forever. However, because the beliefs evolve, the agent may switch to other actions instead of experimenting forever, meaning that the actual value can be higher, which is captured by the adjustable non-negative extra component $C_{Exp}f(p)$.

The new part of the value function arises with the ability to train. The first linear and increasing part represents the average payoff from training the risky arm forever until the good news arrives and switching to purely using this arm immediately after. If the arm is already good (with probability p), the agent on average gets λ in each instance, hence the present value of it being $\frac{\lambda}{r}$, but pays the cost of training up until the first time T when she observes the good news and stops wasting resources, with $T \sim exp(\lambda)$. As such, the agent expects to pay $E_T[\int_0^T (e^{-rt}\kappa dt)] = \frac{\kappa}{r+\lambda}$. In turn, if the arm is bad (with probability $1-p$), the agent expects to start getting the good arm's payoff only once the training succeeds; denote this time as τ. Obtaining this payoff is discounted by $E_\tau[e^{-r\tau}] = \frac{\pi}{r+\pi}$, since $\tau \sim exp(\pi)$

by construction. Additionally, the cost of training has to be paid in each moment of time in $[0, \tau]$, with the present value of it being $E_\tau[\int_0^\tau (e^{-rt}\kappa dt)] = \frac{\kappa}{r+\pi}$. Overall, the first part highlights that the training mechanism provides a sort of insurance to the agent, where she now gets a positive payoff if the arm is bad but has to pay for it even in the good state, since the arm's type is not revealed immediately with training. In turn, the last part of the value function represents the extra value from switching between options given the potential belief evolution under training the arm, and is assumed to be non-negative by restricting solutions to $C_{Tr} \geq 0$. Having more options cannot hurt the agent, and if she strictly benefits from it, the free constant $C_{Tr} > 0$ will be determined as part of the optimal strategy.

3.5 Benchmark with no possibility to train

A useful benchmark is the case where the option of training the risky arm is unavailable. The following lemma reiterates the result previously established in KRC (see the paper for the proof).

Lemma 2 (Benchmark) *In the absence of training, the optimal solution follows a cutoff structure with belief* $\hat{p} = \frac{sr}{\lambda(\lambda - s + r)} \in (0, 1)$ *such that experimentation occurs for* $p > \hat{p}$*, and the safe arm is used otherwise. The solution is characterized by the continuous, smooth and globally convex value function:* $V(p) = max\{\frac{s}{r}, \frac{\lambda}{r}p + \left(\frac{s}{r} - \frac{\lambda}{r}\hat{p}\right)\frac{f(p)}{f(\hat{p})}\}$*, where* $f(p) = \frac{(1-p)^{1+r/\lambda}}{p^{r/\lambda}}$.

In the optimal solution, the agent experiments only if she is sufficiently optimistic. She engages with the risky arm more than her myopic counterpart (who only cares about the current payoff and quits for $p < \frac{s}{\lambda}$), because of the extra positive value from learning while experimenting and anticipating the future change in her actions that is captured by the $\left(\frac{s}{r} - \frac{\lambda}{r}\hat{p}\right)\frac{f(p)}{f(\hat{p})}$ term in the value function. The agent's belief gradually falls, as long as no news arrives, until she reaches a cutoff where she quits experimentation. At such cutoff $\hat{p}$, the agent's value from the safe and the risky arm match (*value matching* condition). There, she is also indifferent between experimenting for an extra instance and then returning to the

optimal path immediately after and not experimenting at all. This is a marginal incentive which, given that the agent becomes more pessimistic with using the risky arm, translates to the *smooth pasting* condition ($V'(\hat{p}) = 0$).

In what follows I will refer to the cutoff and the value function from Lemma 2 as $\hat{p}, V_{bench}(p)$, respectively.

4 Main results

Being able to train interferes with the incentives to experiment and may significantly change the agent's optimal behaviour. Specifically, introducing the training mechanism of a certain efficiency can result in one of the two distinct regimes. The proposition below states these key results.

Proposition 1 *Assume sufficiently low cost of training.* [5] *Denote as* $\underline{p}$ *and* $\overline{p}$ *the lowest and highest belief, under which the agent trains the risky arm, respectively. Then, there exist* $\underline{p}(\pi, \kappa)$ *and* $\overline{p}(\pi, \kappa)$*, such that:*

i. If $\frac{\pi}{\lambda} < \underline{p}(\pi, \kappa)$*, the agent's strategy is non-monotone in actions and leads to quitting experimentation for any* p_0*, unless news arrives: She uses a safe arm for* $p < \hat{p}$*, trains the risky arm in* $(\underline{p}, \overline{p}) \subset [\hat{p}, 1]$*, and purely experiments otherwise.*

ii. If $\frac{\pi}{\lambda} > \underline{p}(\pi, \kappa)$*, the agent's strategy leads to experimenting 'forever' until news arrives for any* $p_0 \geq \underline{p}$*. She gets stuck at training the arm at full intensity until news arrives if* $\frac{\pi}{\lambda} \in [\underline{p}(\pi, \kappa), \overline{p}(\pi, \kappa)]$*, and at some reduced intensity* $\beta^* \in (0, 1)$ *if* $\frac{\pi}{\lambda} > \overline{p}(\pi, \kappa)$*.*

Under lower efficiency of training (when $\frac{\pi}{\lambda} < \underline{p}(\pi, \kappa)$), the agent trains only in some narrow region and gives up experimentation after not receiving a breakthrough for sufficiently long. That is, the agent mainly uses the risky arm to learn its type, and training enables her to increase the chances of observing success. Interestingly, her strategy is then non-monotone

[5] Specifically, $\kappa < \overline{\kappa}(\pi)$, where $\overline{\kappa}(\pi)$ is defined as a highest cost, at which the agent is willing to train the arm for at least a single belief. If $\kappa \geq \overline{\kappa}(\pi)$, the agent never trains the arm, and her solution matches the benchmark one established in Lemma 2. See precise definition and derivation of $\overline{\kappa}(\pi)$in Appendix B.

in actions: She gives up training and purely experiments before quitting experimentation, just as she does when being very optimistic.

In contrast, for $\frac{\pi}{\lambda} > \underline{p}(\pi, \kappa)$, the agent may get stuck on a certain belief and pursue training at some intensity until the news arrives. She converges to such belief both from above and from below: She trains the arm if she is more pessimistic, which rises her belief, while training the arm or purely experimenting in more optimistic states makes her more pessimistic and reduces her belief. Such strategy implies that she does not value learning per se that much and is mainly motivated by ensuring the arm succeeds when she uses it.

Importantly, the two regimes highlight that the ability to train the risky arm impacts experimentation through two distinct channels. First, training has a direct effect on the chance of successful outcome by turning a bad arm into a good one. Second, it increases the likelihood of learning that the good risky arm is good, an indirect and more subtle benefit. Training counteracts the negative impact of having no news, so the agent at least gets disappointed slower, or possibly even becomes more optimistic with training. This implies that the agent anticipates using the risky arm over a longer horizon than she would in the absence of training, which increases her chance of observing a breakthrough from the good arm before she stops. Clearly, the impact of these two channels varies across the regimes and leads to qualitatively different outcomes.

In the rest of the section I motivate the findings in more detail. Subsections 4.1 and 4.2 span the key results under low and high efficiency training, respectively. I then summarise the findings by discussing under which conditions each regime occurs and outlining the overall implications of training on experimentation in Subsection 4.3.

4.1 Low efficiency of training: Non-monotonicity in actions

The agent's optimal strategy under low training efficiency according to Proposition 1 part i can be represented by a scheme as in Figure 2. If the training mechanism is not too efficient (that is, if $\pi < \lambda \underline{p}(\pi, \kappa)$), training may improve the value of experimenting, but

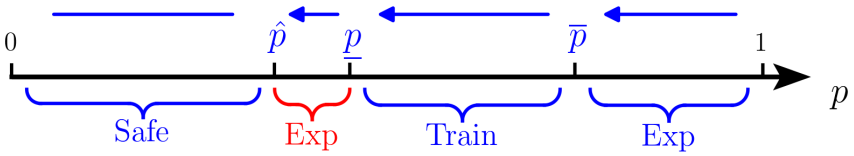

Figure 2: Agent's optimal strategy structure if $\frac{\pi}{\lambda} < \underline{p}(\pi, \kappa)$, as a function of belief p. The arrows above the real line indicate the direction of beliefs evolution in the absence of news.

does not affect the overall strategy qualitatively. In the absence of news, the agent becomes more pessimistic with using the risky arm irrespective of whether she trains it, since she is only willing to train for beliefs above $\frac{\pi}{\lambda}$, where the negative effect from no news dominates the positive influence of training. This implies that, no matter how optimistic the agent is initially, unless the news arrives her belief gradually decreases and converges to $\hat{p}$, where she quits experimentation, as she would without the ability to train.

An interesting feature of this optimal strategy is that the agent's path in the absence of news is necessarily non-monotone in actions. If she starts sufficiently optimistic, she uses the risky arm, then starts training it once she is no longer convinced that the arm is good, and stops training a while before she gives up on the risky arm entirely. That is, she experiments purely for beliefs both higher and lower than the training region, even if the training cost is infinitesimally low. Intuitively, the benefit of training increases when p is lower, since it can only improve the bad arms. As such, its benefit almost disappears when the agent is very optimistic, and training is wasteful then. This rationalizes the agent's incentive to forgo training in $[\overline{p}, 1]$ for any positive cost.

The incentive to give up training but still experiment in $(\hat{p}, \underline{p})$ hinges on the belief dynamic and anticipated action switches in the future. The agent understands that the realization of benefit from training is delayed in time: If she trains the arm today and this training succeeds, it will take some time until the improved arm generates a breakthrough. However, she also anticipates that she will quit experimentation once her belief reaches the stopping

cutoff $\hat{p}$ in some deterministic time, and she will never attain the benefit of training thereafter. Thus, when she is sufficiently close to $\hat{p}$, her marginal benefit from training gets very low, as she is very unlikely to learn about the newly occurring successes of training in the short time left before quitting the risky arm. This makes training no longer worthwhile and she stops training the arm. At the same time, she still has an incentive to experiment for a bit longer, since there is a positive chance of observing the breakthrough if the arm was originally good or the training succeeded in the past. She does so until her belief drops to $\hat{p}$, the benchmark cutoff: Once she quits training, her incentives fully coincide with the ones she would have in the absence of training option, and so she switches to the safe arm at the same belief.

Formally, the incentives to forgo the training are reflected in the marginal benefit of it, $V'(p)\pi(1-p)$, compared to the marginal cost κ. For high beliefs the direct benefit $\pi(1-p)$ is low, while for lower ones $V'(p)$ is low. At $\hat{p}$, $V'(\hat{p}) = 0$, by smooth pasting, as explained in section 3.5. Therefore, the equivalence of the marginal cost and benefit of training must arise for $\underline{p} > \hat{p}$, since $V'(\underline{p}) > 0$ must hold for the benefit to cover even an infinitesimally small cost of training. The higher the cost is, the smaller the training incentives are, so $\underline{p}$ increases with the cost, while $\overline{p}$ declines.

The proposition below characterises the optimal agent's strategy formally, and the remainder of the subsection briefly motivates this technical characterisation.

Proposition 2 *If $\frac{\pi}{\lambda} < \underline{p}(\pi, \kappa)$, the agent's optimal strategy is fully characterized by a continuous, globally convex and smooth value function, with $\underline{p} \in (\frac{\pi}{\lambda}, \underline{p}(\pi, \kappa))$:*

$$V(p) = \begin{cases} \frac{s}{r} & p < \hat{p} \text{ (safe arm)} \\ \frac{\lambda}{r}p + (\frac{s}{r} - \frac{\lambda}{r}\hat{p})\frac{f(p)}{f(\hat{p})} & p \in [\hat{p}, \underline{p}] \text{ (pure exp)} \\ y(p) + \left(\frac{\lambda}{r}\underline{p} - y(\underline{p}) + (\frac{s}{r} - \frac{\lambda}{r}\hat{p})\frac{f(p)}{f(\hat{p})}\right)\frac{g(p)}{g(\underline{p})} & p \in [\underline{p}, \overline{p}] \text{ (training)} \\ \frac{\lambda}{r}p + \left(y(\overline{p}) - \frac{\lambda}{r}\overline{p} + \left(\frac{\lambda}{r}\underline{p} - y(\underline{p}) + (\frac{s}{r} - \frac{\lambda}{r}\hat{p})\frac{f(p)}{f(\hat{p})}\right)\frac{g(\overline{p})}{g(\underline{p})}\right)\frac{f(p)}{f(\overline{p})} & p > \overline{p} \text{ (pure exp)} \end{cases}$$

Notice that the value function that characterizes the low efficiency of training solution

(also plotted in Figure 8 in Appendix A) reflects the anticipated action dynamics. The agent benefits from switching to the other options for lower beliefs, and the non-linear components of $V(p)$ represent these options value. She gains extra value from starting to train the arm below $\overline{p}$, giving up on training at $\underline{p}$, and finally switching to the safe arm at $\hat{p}$ (or a subset of these depending on initial p), and the impact of these switches increases the closer the agent is to making a certain switch. That is, the latter components are decreasing in beliefs (as all the cutoffs are approached from the right) and convex.

Moreover, the value function shares the standard features as in the benchmark outlined in Section 3.5: $V(p)$ is continuous and smooth around all the cutoffs $(\hat{p}, \underline{p}, \overline{p})$. At any cutoff, the agent is indifferent only if she receives the same value from following any of the two actions and if she cannot benefit by marginally deviating to one of the actions for an instance and returning to the optimal path immediately after. The former condition, value matching, guarantees continuity of the value function. The latter translates to smooth pasting at all cutoffs, which hinges on the monotonically decreasing beliefs dynamics. For example, consider the lower boundary for training, $\underline{p}$. Training for an extra instance at $\underline{p}$ makes the agent more pessimistic and pushes her to the left of the cutoff, where she strictly prefers to experiment purely. Hence, she is indifferent in her marginal deviation to training as opposed to immediate pure experimentation only if its marginal cost and benefit conditional on experimenting are equal ($\lim_{p\to\underline{p}_-} V'(p)\pi(1-p) = \kappa$). This results in the smooth pasting requirement, $\lim_{p\to\underline{p}_-} V'(p) = \lim_{p\to\underline{p}_+} V'(p)$, once combined with value matching. A similar logic applies at the other cutoffs.

4.2 High efficiency of training: Convergence to training

Figures 3 and 4 illustrate the possible optimal strategies according to Proposition 1 part *ii*. Strikingly different from traditional Poisson bandits models of experimentation, if the training mechanism is sufficiently efficient, the agent may never give up on the risky arm, even if no breakthrough arrives for a long time. This occurs whenever $\frac{\pi}{\lambda}$ is above $\underline{p}$. At $\underline{p}$

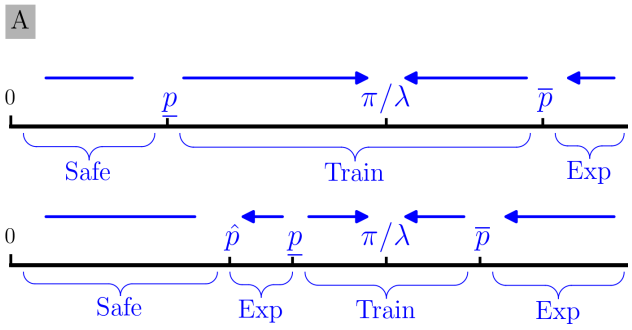

Figure 3: Agent's optimal strategy structure and beliefs evolution for moderately high training efficiency: for lower cost on the top, and higher cost at the bottom.

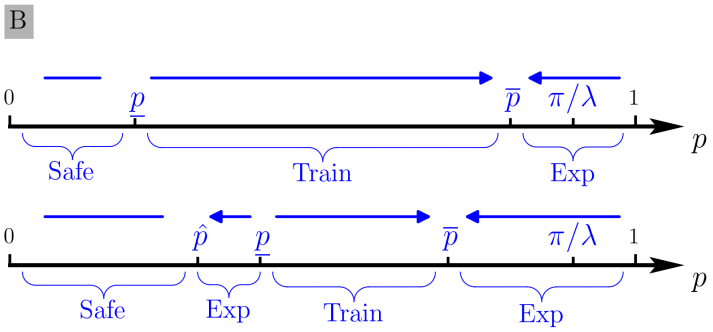

Figure 4: Agent's optimal strategy structure and beliefs evolution for very high training efficiency: for lower cost on the top, and higher cost at the bottom.

the agent becomes more optimistic with training, since its positive effect on belief dominates the negative one of receiving no news, and her belief will keep rising to $\frac{\pi}{\lambda}$ conditional on her pursuing training. If the cost of training is relatively low (when $\frac{\pi}{\lambda} \leq \overline{p}(\pi, \kappa)$), the agent finds it optimal to follow this path and get stuck at belief $\frac{\pi}{\lambda}$, where she keeps training the arm forever until it becomes good and generates a breakthrough. Starting at $p_0 > \frac{\pi}{\lambda}$, the agent gets more pessimistic with purely experimenting for beliefs above $\overline{p}$, switches to training at $\overline{p}$, which slows down her pessimistic learning but does not compensate it fully, so she gradually converges to $\frac{\pi}{\lambda}$ where she again gets stuck at training until the arm succeeds.

As training becomes more costly, her incentive to train at $\frac{\pi}{\lambda}$ disappears. Specifically, when $\frac{\pi}{\lambda} > \overline{p}(\pi, \kappa)$, the marginal benefit of training at $\frac{\pi}{\lambda}$ is lower than its marginal cost, either because the cost κ is too high, or the arm is very unlikely to be bad ($p = \frac{\pi}{\lambda}$ is close to 1), so the direct benefit of training almost vanishes. This implies that the agent quits training for beliefs around $\frac{\pi}{\lambda}$ and only experiments, which leads to her getting more pessimistic in the absence of news. Therefore, the agent can never achieve high enough belief through training and instead converges to the cutoff $\overline{p}$ both from below if she is more pessimistic and trains the arm, and from above if she is more optimistic and experiments purely. At such belief $\overline{p}$, she is exactly indifferent between training the arm and not, and there exists a unique stationary training intensity $\beta^* \equiv \frac{\lambda}{\pi}\overline{p} \in (0, 1)$ (given that $\overline{p} < \frac{\pi}{\lambda}$), such that she optimally

remains at this boundary and trains the arm at reduced intensity until it succeeds.

In both of the scenarios outlined above, the agent ends up experimenting forever for any $p_0 \geq \underline{p}$, by converging to an interior belief and getting stuck there until the breakthrough arrives (due to training). Yet, if she is slightly more pessimistic than $\underline{p}$, she may still have some incentives to experiment: She may prefer to experiment without training for beliefs in $(\hat{p}, \underline{p})$, which is non-empty under high training costs. This creates a sharp discontinuity in the expected outcomes of experimentation around the cutoff $\underline{p}$, since experimenting below $\underline{p}$ eventually dies out unless the news arrives, in contrast to the experimentation above $\underline{p}$. I discuss the welfare implications of this observation further in Section 5.1.

Here, the motivation of forgoing training in $(\hat{p}, \underline{p})$ is different from the one in the low efficiency case, discussed in Section 4.1. Since the agent is willing to train at $\underline{p}$, the marginal benefit of training there (taking into account that she will keep training forever if she starts) should at least weakly exceed its cost. Starting training at $\underline{p} - \epsilon$ (for $\epsilon \to 0$) with continuing doing so forever after incurs a larger marginal benefit, as the arm is slightly more likely to be bad and thus gain from training. However, it also incurs the opportunity cost that training forever gives rise to, as the agent foregoes the safe arm's payoff forever the instance she starts training. This opportunity cost, relative to the benefit of training until news arrives, gets larger the more pessimistic the agent is.

If the cost of training is high, the relative opportunity cost becomes very large sooner (for higher belief), as the benefit of training the arm until the news arrives sharply decreases with the direct training cost. That is, the agent may prefer to quit training for some belief above $\hat{p}$, where she still values learning about the risky arm, but would rather get a safe payoff in the absence of news than train the arm over a long horizon. Hence, she purely experiments in $(\hat{p}, \underline{p})$. Formally, while the benefit of training must at least weakly exceed the cost for $p > \underline{p}$ ($\lim_{p \to \underline{p}_+} V'(p)\pi(1-p) \geq \kappa$), $\lim_{p \to \underline{p}_-} V'(p)\pi(1-p) \leq \kappa$ may hold, given that the agent anticipates different action dynamics on the different sides of the cutoff. This ultimately guarantees that for any belief below $\underline{p}$, the agent does not benefit from training for

a short instance, so can never become optimistic enough to reach $\underline{p}$ and commence training the risky arm forever.

As the costs get lower, the value of training the risky arm until it succeeds increases for any belief, broadening the incentives to train (i.e. lowering $\underline{p}$). Once the cost is low enough, the value of training forever starting from belief $p_0 = \hat{p}$ strictly exceeds the value of marginally experimenting and giving up, implying that at the lower training cutoff the agent is indifferent between training forever and sticking to the safe arm immediately, $\underline{p} < \hat{p}$. Such cutoff decreases further with the cost, and if the reduction in cost is coupled with especially high training productivity π, it is possible that the agent is willing to train even a certainly bad risky arm, i.e. at $p = 0$, implying that any risky arm will be experimented upon forever.

The proposition below provides a technical characterization of the solution under high efficiency training, and the remainder of the subsection spans some formal aspects of it.

Proposition 3 *If $\frac{\pi}{\lambda} > \underline{p}(\pi, \kappa)$, the agent's optimal strategy is fully characterized by a continuous, globally convex and kinked at $\underline{p}$ (and smooth otherwise) value function.*

i. If $\frac{\pi}{\lambda} \in [\underline{p}(\pi, \kappa), \overline{p}(\pi, \kappa)]$, then $\{\underline{p}, \overline{p}\} = \{\underline{p}(\pi, \kappa), \overline{p}(\pi, \kappa)\}$, and

$$V(p) = \begin{cases} \max\{\frac{s}{r}, \frac{\lambda}{r}p + (\frac{s}{r} - \frac{\lambda}{r}\hat{p})\frac{f(p)}{f(\hat{p})}\} & p < \underline{p} \text{ (safe arm or pure exp)} \\ y(p) & p \in [\underline{p}, \overline{p}] \text{ (training)} \\ \frac{\lambda}{r}p + \left(y(\overline{p}) - \frac{\lambda}{r}\overline{p}\right)\frac{f(p)}{f(\overline{p})} & p > \overline{p} \text{ (pure exp)} \end{cases}$$

ii. If $\frac{\pi}{\lambda} > \overline{p}(\pi, \kappa)$, then $\overline{p} \in (\overline{p}(\pi, \kappa), \frac{\pi}{\lambda})$, and

$$V(p) = \begin{cases} \max\{\frac{s}{r}, \frac{\lambda}{r}p + (\frac{s}{r} - \frac{\lambda}{r}\hat{p})\frac{f(p)}{f(\hat{p})}\} & p < \underline{p} \text{ (safe arm or pure exp)} \\ y(p) + \left(\frac{\lambda}{r}\overline{p} - y(\overline{p}) + \frac{(\lambda\pi(1-\overline{p}) - \kappa r)\overline{p}}{r\pi(\overline{p} + \frac{r}{\lambda})}\right)\frac{g(p)}{g(\overline{p})} & p \in [\underline{p}, \overline{p}] \text{ (training)} \\ \frac{\lambda}{r}p + \left(\frac{(\lambda\pi(1-\overline{p}) - \kappa r)\overline{p}}{r\pi(\overline{p} + \frac{r}{\lambda})}\right)\frac{f(p)}{f(\overline{p})} & p > \overline{p} \text{ (pure exp)} \end{cases}$$

As in Section 4.1, the option value parts of the function reflect the anticipated action

switches given the belief dynamics. Notably, whenever the agent converges to training at full intensity at $\frac{\pi}{\lambda}$ (for $\frac{\pi}{\lambda} \in [\underline{p}(\pi,\kappa), \overline{p}(\pi,\kappa)]$), once she starts training, she never switches to other options (until the news arrives), and so the value function $V(p)$ is linear in the training region and contains only the value of training the arm forever, $y(p)$ (see figure 9 in Appendix A). Experimenting above $\overline{p}$ adds some value to the agent because she ends up stuck at training as well, which is reflected in the option value part.

As the cost rises to $\frac{\pi}{\lambda} > \overline{p}(\pi,\kappa)$, the agent converges to training the arm at a reduced intensity β^* and gets the extra value from it both when experimenting in $(\overline{p}, 1)$ and when training in $(\underline{p}, \overline{p})$. Indeed, while training for $p < \overline{p}$, she anticipates getting stuck at $\overline{p}$ and hence saving on the training costs when they would have exceeded the benefit - at $\overline{p}$, $V'(p)\pi(1-p) = \kappa$ exactly. Similarly, while using the arm for $p > \overline{p}$, she expects to start training at intensity β^* instead of becoming pessimistic with using the arm forever. Following the above, the options value components in the value function for $p > \underline{p}$ are increasing towards $\overline{p}$ (increasing and convex in $(\underline{p}, \overline{p})$, and decreasing and convex in $(\overline{p}, 1)$) (see figure 10 in Appendix A).

Under both scenarios, the optimal action profile is sharply divided by the lower cutoff $\underline{p}$, where being above $\underline{p}$, the agent can never become more pessimistic than $\underline{p}$, and vice versa. As such, the agent's value above $\underline{p}$ does not depend on anything occurring below $\underline{p}$, including possibly experimenting there or using the safe arm, as well as the value of the cutoff itself. Similarly, being below $\underline{p}$, the agent cannot become more optimistic than $\underline{p}$, so her value for $p < \underline{p}$ is independent of $\underline{p}$ and anything above this cutoff.

The diverging behaviour around the lower training cutoff $\underline{p}$ implies that its value is determined fully by the value matching requirement, and it necessarily violates the smooth pasting as a result. Since training makes the agent more optimistic at $\underline{p}$, a short (marginal) deviation to training with returning to the optimal path immediately after pushes her deeper to the training region, implying that the value from such deviation fully matches the value to the right of the cutoff. Similarly, a short deviation to the action to the left of $\underline{p}$ (either using a safe arm or purely experimenting) either leaves the agent's belief unchanged or pushes her

more to the left, thus matching $\lim_{p\to\underline{p}_-} V(p)$. Hence, any marginal indifference condition coincides with the value matching one, and smooth pasting that typically arises due to this marginal incentive need not hold.

When $\underline{p}$ arises from indifference between training the risky arm forever and using the safe arm (if $\underline{p} < \hat{p}$), the marginal cost of training, $s - \lambda\underline{p} + \kappa$, is larger than the benchmark one, $s - \lambda\hat{p}$, while the jump benefit from learning the news, $\lambda\underline{p}\frac{\lambda - s}{r}$, is smaller. So the marginal cost and benefit of training at $\underline{p}$ can be equal only if the smooth benefit under no news is strictly positive; i.e. $\lim_{p\to\underline{p}_+} V'(p) > 0$ must hold, violating smooth pasting. If the agent is indifferent between experimenting with and without training at $\underline{p}$, we can show that the marginal benefit of training exceeds the marginal cost to the right, $\lim_{p\to\underline{p}_+} V'(p)\pi(1-p) > \kappa$, while the cost dominates to the left, $\lim_{p\to\underline{p}_-} V'(p)\pi(1-p) \leq \kappa$, guaranteeing the convex kink and smooth pasting violation as well. [6]

Note that the value function remains smooth at the upper cutoff $\overline{p}$, where the agent switches from using the risky arm to training it. If the agent marginally forgoes training at $\overline{p}$, she becomes more pessimistic under pure experimentation and thus switches to training immediately. She is indifferent to such deviation only if the forgone cost of training matches the forgone training benefit, $V'(\overline{p})\pi(1-\overline{p}) = \kappa$, which, together with value matching, guarantees smooth pasting. Intuitively, since the threshold is absorbing at least from the right, if the agent approaches it from there, she will purely experiment for as long as the marginal cost of training exceeds its benefit and switch to improving the arm as soon as the benefit exceeds the cost; such strategy guarantees a smooth transition from one action to another.

If $\frac{\pi}{\lambda} > \overline{p}(\pi, \kappa)$, the upper cutoff $\overline{p}$ is absorbing on both sides - this is a novel boundary type, that typically does not occur in experimentation models. To define such cutoff, the traditional conditions of value matching and smooth pasting do not suffice, because they only identify two of the three free parameters in the triplet $\{\overline{p}, C_{Tr}, C_{Exp}\}$ (i.e. define

[6] Such 'irregularity' of the cutoff $\underline{p}$ is similar to the one identified in Keller and Rady (2015) for the bad news information structure. However, the reason for the smooth pasting violation is different here. In KR it relies on the disappearance of the jump benefit at the cutoff (as learning the bad news result in quitting the experimentation, which coincides with the cutoff payoff exactly).

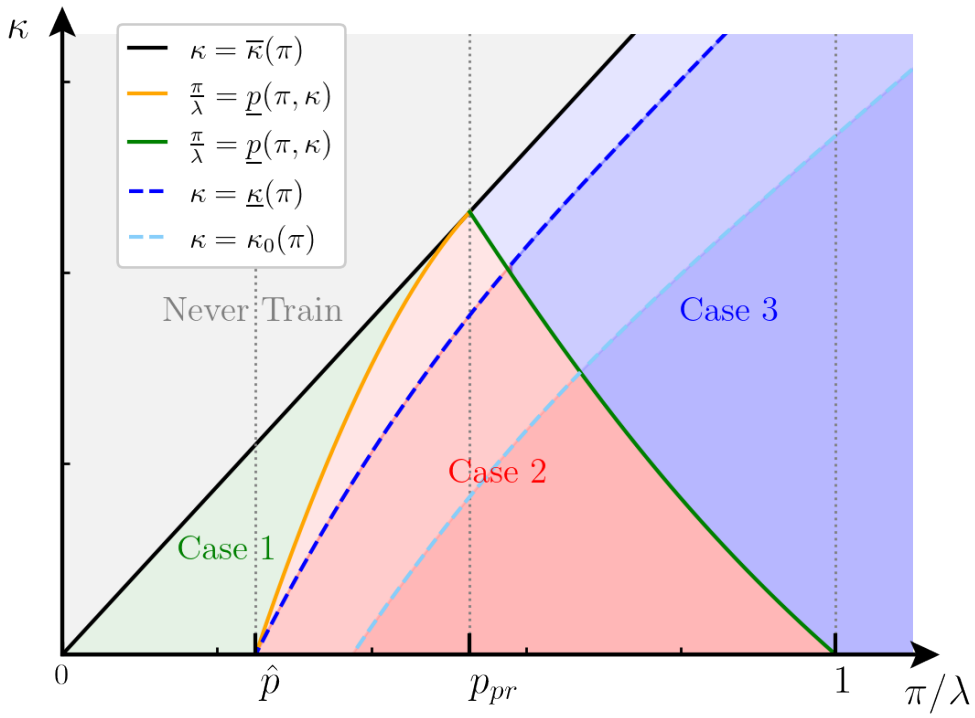

Figure 5: Training regimes for all possible values of π, κ, with π normalized to $\frac{\pi}{\lambda}$. $\frac{\pi}{\lambda} = \underline{p}(\pi, \kappa)$ and $\frac{\pi}{\lambda} = \overline{p}(\pi, \kappa)$ are defined in Proposition 1 and separate the space into three cases: case 1 - low efficiency training, case 2 - high efficiency with convergence to $\frac{\pi}{\lambda}$, case 3 - high efficiency with convergence to $\overline{p}$. The rest are informally defined and discussed in subsection 4.3.

$C_{Tr}(\overline{p}), C_{Exp}(\overline{p})$). The extra necessary condition guarantees that the agent cannot improve her value by shifting the cutoff of interest slightly, that is $\overline{p} \equiv \arg\max_q V(p; q)$, where $V(p; q) = \frac{\lambda}{r}p + C_{Exp}(q)f(p)$, implying that the boundary is defined through $C'_{Exp}(\overline{p}) = 0$ (or $C'_{Tr}(\overline{p}) = 0$, equivalently). If this extra optimality condition holds, the agent is indifferent between training only at the cutoff $\overline{p}$, strictly prefers training to the left of the cutoff and pure experimentation to the right, and cannot benefit from changing the intensity of training anywhere. [7]

4.3 Key findings summary and implications

Overall, there are two regimes arising from the ability to train: The experimentation incentives dominate in one, while the motive of improving the arm is prevailing in the other. Figure 5 illustrates when one or the other case arises for any training mechanism in the studied class, i.e. for any combination of π and κ, keeping the baseline parameters, $\{\lambda, s, r\}$, fixed. Training can only benefit the agent if it is not too costly, so she never trains the arm in the top left corner of the diagram. Then, $\frac{\pi}{\lambda}$ vs $\underline{p}(\pi, \kappa)$ constraint determines which of the two regimes occurs. The agent predominantly experiments with only marginal impact of training in the region in the bottom left (Case 1 region), which I previously referred to as 'low training efficiency' and studied in 4.1: That is, when the training productivity π is low, while its cost κ is moderately high. Otherwise, when training is either cheaper or more productive (to the right of the $\frac{\pi}{\lambda} = \underline{p}(\pi, \kappa)$ constraint), it has a more salient effect and changes the set of convergence outcomes, as described in section 4.2. Within this high efficiency region, the agent can either converge to training at full intensity at $p = \frac{\pi}{\lambda}$ (Case 2 on the diagram) or at reduced intensity at $p = \overline{p}$ (Case 3), which are separated along the $\frac{\pi}{\lambda} = \overline{p}(\pi, \kappa)$ constraint. Intuitively, the agent is willing to save some training costs by reducing intensity if the costs are high, so the latter scenario occurs above the constraint.

Being able to train not only improves the outcome of a bad risky arm, it also increases the likelihood of observing a breakthrough from initially good arms before quitting experimentation. The proposition below breaks down the qualitative impact of these effects on experimentation.

Proposition 4 *Having an ability to train:*

i. Improves the value of experimenting (for $p > \underline{p}$) if $\kappa < \overline{\kappa}(\pi)$, where $\overline{\kappa}(\pi)$ is a cost at which the agent is willing to train only at a single belief.

[7] To confirm that there are no further improvements, define $V_{indiff}(q) \equiv V(q; q)$ - such value function represents the value of being indifferent between experimenting with and without training and thus training at any intensity $\beta \in [0, 1]$; such $V_{indiff}(q)$ is strictly concave, and plotted in Figure 10. Then, the optimality condition $C'_{Exp}(\overline{p}) = 0$ implies $V'_{indiff}(\overline{p}) = V'(\overline{p})$, and so $V(p) \geq V_{indiff}(p)$ for any p, with equality occurring only at $p = \overline{p}$.

ii. Broadens the incentives to experiment beyond $\hat{p}$ if $\kappa < \underline{\kappa}(\pi)$, where $\underline{\kappa}(\pi)$ is a cost at which $\underline{p} = \hat{p}$. If $\frac{\pi}{\lambda} \leq \hat{p}$, this effect never occurs, even if training is costless.

iii. Guarantees a breakthrough almost surely for all prior beliefs above $\underline{p}$ if $\frac{\pi}{\lambda} > \underline{p}(\pi, \kappa)$.

The qualitative impact of training on experimentation varies across the training regimes. Trivially, if the agent trains the risky arm or anticipates doing so in the future, it improves her continuation value, otherwise she would not engage with the training mechanism. This occurs only if the cost is sufficiently low ($\kappa < \overline{\kappa}(\pi)$), as for higher costs training is fully dominated by pure experimentation, as indicated on Figure 5. Under lower efficiency (when $\frac{\pi}{\lambda} < \underline{p}(\pi, \kappa)$), this is the only impact training has on the agent. However, for $\frac{\pi}{\lambda} > \underline{p}(\pi, \kappa)$, the ability to train causes a change to the experimentation incentives and outcomes. Not only it enhances the value of experimenting, it also increases the incentives to experiment beyond $\hat{p}$ under low enough costs ($\kappa < \underline{\kappa}(\pi)$), and ensures that the breakthrough arises with probability 1 almost surely for priors above $\underline{p}$.

Intuitively, the effect of increasing the scope of experimentation to a wider range of beliefs should occur for any training productivity under low enough costs. This is, however, not so for low training productivity levels $\pi \leq \lambda\hat{p}$: The agent does not experiment more even if $\kappa = 0$ then. Under costless training, the only effect the training has as opposed to pure experimentation is in the marginal beliefs shift conditional on obtaining no news (this effect is weakly positive, so training dominates over pure experimentation). If the training efficiency is low, the agent does not benefit from this effect at the stopping cutoff, because she gets pessimistic there and smoothly transitions to the safe option (by Proposition 1, smooth pasting holds at $\underline{p}$). As a result, her incentive to train fully coincides with her incentive to experiment in the absence of training, and thus, $\underline{p} = \hat{p}$. Note that having a training option broadens the agent's incentives to experiment under low enough costs (when $\underline{p} < \hat{p}$), but can never reduce those (safe arm is never used for $p > \hat{p}$): If the agent does not want to train in the neighbourhood of $\hat{p}$, she still benefits from learning the risky arm's type there, and hence purely experiments at least.

Finally, the agent obtains a breakthrough almost surely whenever she gets stuck at training the arm with a least some intensity until the news arrives, i.e. if $\frac{\pi}{\lambda} > \underline{p}(\pi, \kappa)$ holds (implying $\frac{\pi}{\lambda} > \underline{p}$). Such dynamic ensures that she will keep experimenting 'forever', and if the arm is initially good, she will learn it eventually, $\lim_{\tau \to \infty} \Pr(\text{news by } \tau | \lambda^\theta = \lambda) = 1$. That is, training strengthens experimentation incentives so much that the agent never mistakenly abandons a good risky arm, which can occur with positive probability in the absence of training option. Similarly, if the arm is initially bad, the agent will train it 'forever' until it becomes good and generates a breakthrough, which also occurs almost surely in the limit.

There is a variation in the range of the priors which lead to a guaranteed breakthrough within the region of $\frac{\pi}{\lambda} > \underline{p}(\pi, \kappa)$, and Figure 5 reflects that. For low costs and high enough training productivity, the agent trains even a surely bad risky arm, so any prior belief of the agent results in a breakthrough. This region is constrained from above by $\kappa = \kappa_0(\pi)$, which indicates the highest cost at which $\underline{p} = 0$. The training incentive gradually narrows for costs in $(\kappa_0(\pi), \underline{\kappa}(\pi))$, so the agent never touches a risky arm if she is very pessimistic about it, but otherwise takes it and almost surely turns it to a breakthrough. If the cost gets even higher ($\kappa > \underline{\kappa}(\pi)$), an extra pure experimentation region, $(\hat{p}, \underline{p})$, arises, and the instances at which the agent necessarily obtains a breakthrough are only a subset of those where she is willing to experiment: Any experimentation below $\underline{p}$ will eventually be terminated in the absence of news.

5 Discussion and extensions

In this section I build on the key results to draw further implications of training on experimentation. I start by presenting a way to think about welfare in this setup in subsection 5.1. I then move to extending results for the information structure containing bad news and to a convex cost of training in subsections 5.2 and **??**, respectively. I overall show that the finding of being able to pursue experimentation 'forever' preserves under these extended

setups, but possibly obtained differently.

5.1 Welfare implications

Clearly, the agent's preference of one experimentation regime over another transitions smoothly. That is, the agent never has a discontinuous increase in her continuation value from a marginal change in any parameter. Yet, there is potential discontinuity in the expected outcome of experimentation that arises due to being able to train. Recall that under high efficiency of training and high enough cost the agent prefers experimenting both above and below the lower training cutoff $\underline{p}$. If she is above, she will keep training the arm until it generates a breakthrough, which occurs with probability 1 almost surely in the limit. In contrast, being below, she will use the risky arm for a short while to learn if it is good and abandon it in a deterministic time it takes to move from p_0 to $\hat{p}$ in the absence of news. Such dynamics results in a strictly positive likelihood of ending up with no breakthrough: The bad arms will never generate a success, and some of the good ones are abandoned before their true type is observed.

While in traditional models of experimentation any small change can only marginally impact the likelihood of learning the good news before switching to the safe arm, the above suggests that here it may lead to a discrete jump in the probability of observing success from strictly interior to 1. This discontinuity of the outcome is leveled out for the agent given her discounting, but may matter significantly more generally. An example of such is provided below.

Corollary 1 *Assume a social planner who values the future more than an experimenting agent, and can intervene in designing the prior information. Then, whenever* $\frac{\pi}{\lambda} > \underline{p}(\pi, \kappa)$*, the social planner:*

i. Can achieve a jump gain from marginal intervention in the prior p_0 *around* $\underline{p}$*.*

ii. May strictly prefer incomplete information environment to full information under sufficiently optimistic prior ($p_0 > \underline{p}$) *and sufficiently high training cost* $\kappa > \kappa_0(\pi)$*.*

To illustrate, imagine a social planner (he) who values only the long-term outcome of experimentation and does not care about how the steady state is achieved. Assume he can intervene in the prior information of the experimenting agent (she) through moving her prior from p_0 to some p_0' by paying a cost $c(|p_0' - p_0|)$, with some smoothly increasing $c(\cdot)$ such that $c(0) = 0$. Then, the first observation follows directly: If $\frac{\pi}{\lambda} > \underline{p}(\pi, \kappa)$, experimentation leads to a breakthrough almost surely for any prior above $\underline{p}$, and stops before achieving success with strictly positive probability otherwise. If the prior p_0 is just below the cutoff belief $\underline{p}$ (eg. $p_0 = \underline{p} - \epsilon$ with $\epsilon \to 0$), a slight manipulation that pushes the prior upwards is always attractive, as it gives the social planner a jump gain at a marginal cost of $c(\epsilon) \to 0$.

Interestingly, depending on the prior, the social planner may also strictly prefer to commit to not revealing the information about the arm's type to the experimenting agent. With full information, if $\kappa > \kappa_0(\pi)$, the agent would immediately adopt a good risky arm, but abandon a bad one in favor of the safe option, as training the bad risky arm is too costly for her. In contrast, without the information and with sufficiently optimistic prior ($p_0 > \underline{p}$), she would train the arm and achieve a success almost surely independently of the arm's initial type. This ensures a better outcome and is thus preferred by the social planner.

5.2 Information structure with breakdowns

Model

Consider a flipped information structure - that is, the news arrival is actually a breakdown instead of breakthrough, so the news signals a bad risky arm instead of the good one. I redefine my model to match the one by Keller, Rady (2015). More specifically, suppose a risky arm is of type $\lambda^b \in \{\lambda, 0\}$, with $\Pr(\lambda^b = \lambda) = q_0$. The arm generates a *cost* of size 1 at a Poisson arrival rate λ^b, that is the arm is now 'bad' if $\lambda^b = \lambda$ (it occasionally results in costs, while the 'good' arm never does with $\lambda^b = \lambda$). The safe arm, in turn, yields a guaranteed fixed *cost* $s \in (0, \lambda)$.

Similarly to the main model, the agent can invest in improving the risky arm, and the

'training' mechanism of doing so is exactly as in the breakthroughs model: that is, at a cost κdt paid in dt, the bad risky arm becomes good with probability πdt (i.e. at arrival rate π). As before, once an improvement has occurred, it is irreversible. Note that the instance of the investment's success in not directly observable, and can only be inferred from the absence of breakdowns - that is, given the changed signal structure, the agent is never perfectly sure that the training succeeded (as he is never certain whether the arm is good).

Given that that the payoffs are now in terms of costs as opposed to positive payoffs, the agent is looking to *minimize* the present value of the costs inflow. In particular, in each moment in time he still picks the optimal intensity of experimentation α_t and the optimal intensity of training conditional on experimentation β_t that solve:

$$\min_{((\alpha_t,\beta_t)\in[0,1]^2)_0^\infty} E_{q_0}\left[\int_0^\infty e^{-rt}\left((1-\alpha_t)s+\alpha_t\left(q_t\lambda+\beta_t\kappa\right)\right)dt\right].$$

Beliefs evolution

The following lemma states the evolution of beliefs given a changed information structure.

Lemma

For any q_t and $\alpha_t > 0$, the risky arm is believed to be certainly bad conditional on a cost of size 1 generated by it: $q_{t+dt}^{news} = 1$. Conditional on no cost by risky arm generated, beliefs follow: $dq_t = -\alpha_t q_t(\beta_t\pi + \lambda(1-q_t))dt$.

Notice that conditional on no news arriving the agent surely gets more optimistic (i.e. the belief that the arm is bad, q_t, decreases). This happens because with new modification both the learning effect and the improvement effect work in the same direction: no news is a good news now, and improving the arm boosts the optimism even further. This observation simplifies the analysis a lot: in particular, the main feature of the Poisson bandits preserves - the belief dynamics conditional on no news is monotonic, so unless the agent exploits the safe arm, he converges to believing the risky arm is good, $q = 0$. As such, shifting the information structure from good to bad news makes the model more unified with the

traditional experimentation models, and shuts down all of the novel scenarios that occurred in the model with breakthroughs, as discussed in the main section.

Benchmark - no training

In the absence of the training option, the model fully replicates the one in Keller, Rady (2015). The lemma below outlines their finding.

Lemma (benchmark)

There exists a cutoff $\hat{q}$, such that the agent experiments with the risky arm for $q < \hat{q}$ and uses the safe arm otherwise. The solution is characterised by a piece-wise linear, weakly increasing value function with a concave kink at $\hat{q}$: $V(q) = min\{\frac{s}{r}, \frac{\lambda}{\lambda+r}\left(\frac{s}{r}+1\right)q\}$.

The solution exhibits a cutoff structure, where the agents stops experimentation for high beliefs q and experiments otherwise. This implies that once the agent starts experimenting, conditional on receiving no news, he gradually becomes more optimistic, and his belief moves away from the cutoff - that is, once the agent starts using the risky arm, he will use it forever unless he learns a bad news. Such strategy implies that the only payoff the agent receives from experimentation is the one from experimenting forever and stopping only if he learns the risky arm is bad, which, similarly to the derivations in the main model, is characterized by a linear function. The agent never switches to the other options, hence, the value associated with this (the second non-linear part in the value function) is normalized to 0.

As a result of the above, the belief that makes the agent indifferent between experimenting and not is characterized solely by value matching, and necessarily violates the smooth pasting requirement. At the cutoff, the loss from experimenting $s - \lambda\hat{q}$ is strictly positive, while the jump benefit from learning the news is zero ($V(1) - V(\hat{q}) = \frac{s}{r} - \frac{s}{r} = 0$). Hence, the agent is only indifferent at the cutoff if the smooth benefit from experimentation covers the loss, implying that $\lim_{q\to\hat{q}_-} V'(q) > 0$ - a direct violation of smooth pasting.

I will refer to the solution above as a benchmark and $V_{bench}(q)$ in what follows.

Optimal strategy

The proposition below outlines the optimal strategies for all the possible characteristics

of the training mechanism π, κ.

Proposition 3

1. If $\kappa \geq \frac{s}{r}\pi$, the agent never engages in training; thus, the optimal strategy is the benchmark one characterized by $V_{bench}(q)$.

2. If $\kappa < \frac{s}{r}\pi$, there exist cutoffs $\underline{q} \in [0, \hat{q})$ and $\overline{q} \in (\hat{q}, 1]$, such that the agent optimally improves the risky arm for $q \in [\underline{q}, \overline{q}]$, purely experiments for $q < \underline{q}$, and uses the safe arm for $q > \overline{q}$. The resulting value function is weakly increasing, globally concave, and smooth everywhere except for a single kink at $\overline{q}$, and is characterized by:

$$V(q) = \begin{cases} \frac{\lambda}{\lambda+r}\left(V(1)+1\right)q & q < \underline{q} \\ y(q) - \left(y(\underline{q}) - \frac{\lambda}{\lambda+r}\left(V(1)+1\right)\underline{q}\right)\frac{g(q)}{g(\underline{q})} & q \in [\underline{q}, \overline{q}] \\ \frac{s}{r} & q > \overline{q} \end{cases}$$

where $y(q) = \frac{\lambda}{\lambda+\pi+r}\left(V(1)+1\right)q + \frac{\kappa}{r}(1 - \frac{\lambda}{\lambda+\pi+r}q)$, *and* $g(q) = \frac{(\frac{\pi}{\lambda}+1-q)^{1+\frac{r}{\lambda+\pi}}}{q^{\frac{r}{\lambda+\pi}}}$.

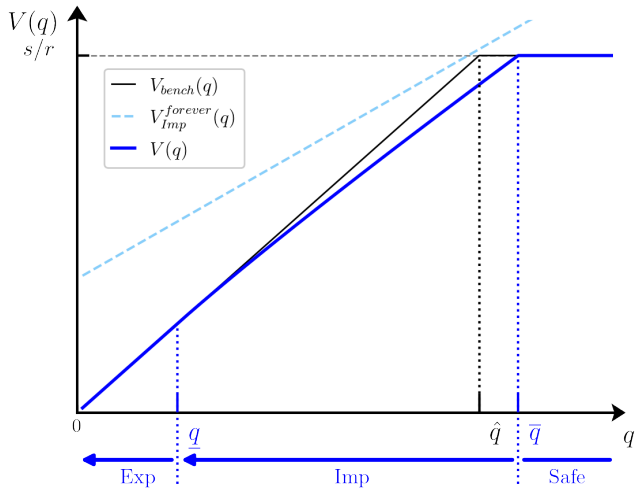

Figure 6: Text.

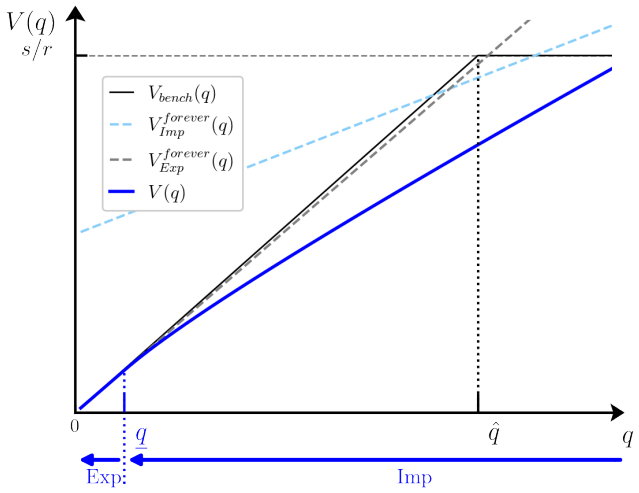

Figure 7: Text.

Similarly to the good news case, the agent trains the risky arm only in some moderate range of beliefs $(\underline{q}, \overline{q})$. If she is sufficiently optimistic $(q < \underline{q})$, she quits training and experiments purely, since the training benefit practically disappears if the arm is very likely

to be good. Similarly, if the arm is very likely bad ($q > \overline{q}$), the agent may prefer to quit experimentation entirely (as plotted in Figure 6). However, if the training mechanism is efficient (in a sense of high productivity π and low cost κ), the stopping region may shrink entirely, with $V(1)$ becoming the value from training the bad arm and switching to pure experimentation at some point in case no bad news arrives. This is plotted in Figure 7.

Notice that the value function $V(q)$ has a concave shape as opposed to linear, which exactly accounts for the agent's anticipation of quitting training and thus saving on the costs after $\underline{q}$. While $V(q)$ remains kinked at $\overline{q}$ (for the same reason as a kink at $\hat{q}$ in the benchmark appears), smooth pasting holds at the lower cutoff $\underline{q}$. Intuitively, the agent becomes more optimistic around this boundary (q reduces), and thus quits training only once she does not want to train even for a single instance, ensuring a smooth transition.

Overall, despite the non-monotonicity in beliefs and actions disappears, the introduction of training in the information structure with breakdowns yields similar qualitative predictions as the main good news model does. Firstly, the breakdowns model necessarily results in both benefits as specified in the breakthroughs environment. Training improves the value conditional on experimenting (for $q > \underline{q}$), and it necessarily broadens the scope for experimentation ($\overline{q} > \hat{q}$ according to Proposition 3).

Secondly, the feature of experimenting forever preserves for sufficiently efficient training mechanisms, just is a different manner rather than being stuck on a certain belief. Indeed, when $\overline{q} = 1$, learning the bad news does not discourage the agent enough to give up the risky arm, as training it gives her a higher payoff than the safe option. This implies that the agent keeps experimenting forever, possibly alternating between training and not (if the news does not arrive for sufficiently long). As in the main breakthroughs model, such strategy guarantees that the risky arm becomes good with probability 1 in the limit. Otherwise, under weaker training mechanisms, the agent gives up after the first breakdown, as she would in the absence of training opportunity.

6 Concluding remarks

Overall, the ability to train risky arms enriches experimentation framework by allowing for various actions trajectories and qualitatively different outcome predictions. Specifically, training affects the possible belief dynamics and makes it non-monotone even in the absence of news. Such observation gives rise to two key results. First, the agent may behave in a non-monotone manner: She quits training the arm a while before giving up on experimentation. Second, the agent may optimally get stuck on a certain interior belief, with converging towards this belief in the absence of news both when being more optimistic by purely experimenting, and pessimistic, by training the arm. In such a strategy, the agent pursues training until the arm generates a success, and so obtains a breakthrough almost surely in the limit. Hence, the ability to train the arm may not just improve the bad arm's performance, but has a spillover effect on the rate of discovering the a priori good arms: Under sufficiently efficient training mechanism all the good risky arms will be discovered and used in the limit, while they are wrongly abandoned with some positive probability in traditional experimentation models.

Importantly, being able to improve the arms challenges our understanding of learning from experimentation. Specifically, I highlight that the agent can reach a success in experimentation by either discovering initially good arms or investing resources in the bad ones for so long that they generate an equal success. This suggests that observing experimentation outcomes of others conveys much less information than typically assumed. As such, this paper opens opportunities for future research on the issues of learning about the options while also improving those, and I hope that bringing this agenda to the models of strategic experimentation among multiple agents may unveil new insights in our understanding of these class of problems.

References

Bergemann, Dirk and Ulrich Hege (1998) "Venture capital financing, moral hazard, and learning," *Journal of Banking Finance*, 22 (6), 703–735.

——— (2005) "The financing of innovation: Learning and stopping," *RAND Journal of Economics*, 719–752.

Bonatti, Alessandro and Johannes Hörner (2011) "Collaborating," *American Economic Review*, 101 (2), 632–663.

Fershtman, Daniel and Alessandro Pavan (2023) "Searching for "Arms": Experimentation with Endogenous Consideration Sets," *Working paper*, `https://sites.google.com/site/danielfershtman`.

Fryer, Roland and Philipp Harms (2018) "Two-armed restless bandits with imperfect information: Stochastic control and indexability," *Mathematics of Operations Research*, 43 (2), 399–427.

Guo, Yingni (2016) "Dynamic delegation of experimentation," *American Economic Review*, 106 (8), 1969–2008.

Halac, Marina, Navin Kartik, and Qingmin Liu (2016) "Optimal contracts for experimentation," *The Review of Economic Studies*, 83 (3), 1040–1091.

——— (2017) "Contests for experimentation," *Journal of Political Economy*, 125 (5), 1523–1569.

Hidir, Sinem (2019) "Contracting for Experimentation and the Value of Bad News," Working paper.

Hörner, Johannes and Larry Samuelson (2013) "Incentives for experimenting agents," *The RAND Journal of Economics*, 44 (4), 632–663.

Keller, Godfrey and Sven Rady (1999) "Optimal Experimentation in a Changing Environment," *The Review of Economic Studies*, 66 (3), 475–507.

——— (2010) "Strategic experimentation with Poisson bandits," *Theoretical Economics*, 5 (2), 275–311.

——— (2015) "Breakdowns," *Theoretical Economics*, 10 (1), 175–202.

Keller, Godfrey, Sven Rady, and Martin Cripps (2005) "Strategic Experimentation with Exponential Bandits," *Econometrica*, 73 (1), 39–68.

Klein, Nicolas and Sven Rady (2011) "Negatively correlated bandits," *The Review of Economic Studies*, 78 (2), 693–732.

Moroni, Sofia (2022) "Experimentation in organizations," *Theoretical Economics*, 17 (3), 1403–1450.

Safronov, Mikhail (2021) "Experimentation and Learning-by-Doing," Working paper.

Whittle, Peter (1988) "Restless bandits: activity allocation in a changing world," *Journal of Applied Probability*, 25 (A), 287–298.

Yariv, Leeat (2021) "Disentangling Exploration from Exploitation," *Proceedings of the 22nd ACM Conference on Economics and Computation*.

A Appendix. Figures

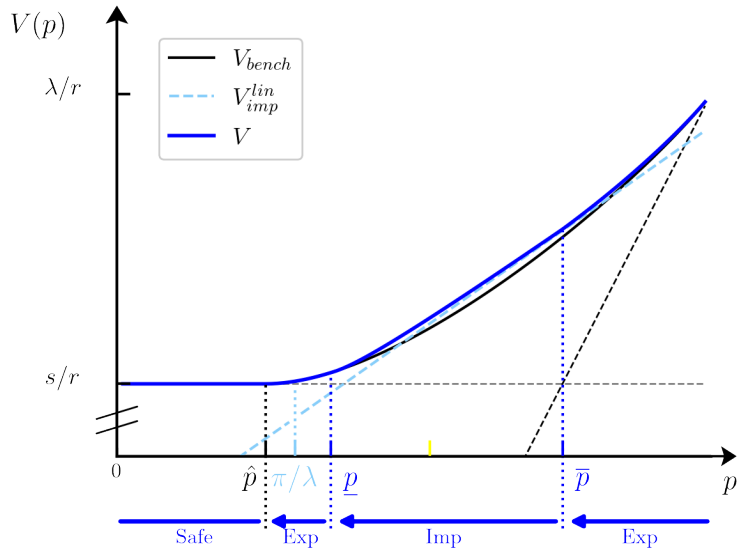

Figure 8: The optimal value function under low efficiency of training ($\frac{\pi}{\lambda} < \underline{p}(\pi, \kappa)$), as defined in Proposition 2.

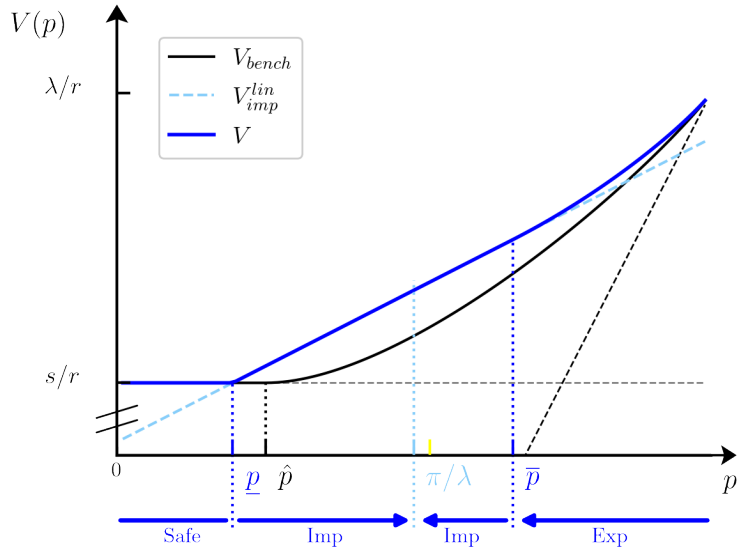

Figure 9: The optimal value function under moderately high efficiency of training ($\frac{\pi}{\lambda} \in [\underline{p}(\pi, \kappa), \overline{p}(\pi, \kappa)]$), as defined in Proposition 3 part *i*. Plotted under relatively low costs, so illustrates the dynamics under top scheme on Figure 3.

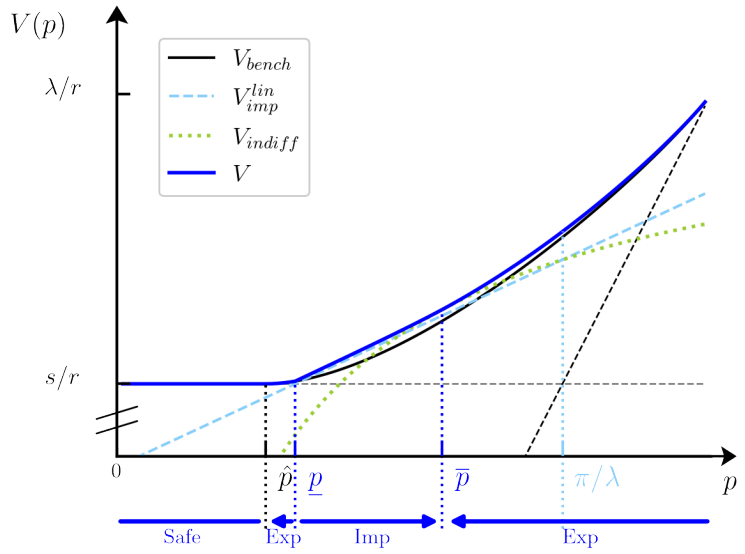

Figure 10: The optimal value function under very high efficiency of training ($\frac{\pi}{\lambda} > \overline{p}(\pi, \kappa)$), as defined in Proposition 3 part *ii*. Plotted under relatively high costs, so illustrates the dynamics under bottom scheme on Figure 4.

B Appendix. Proofs

B.1 Main results (Section 4)

To prove the main results I establish and prove some auxiliary lemmas.

Lemma B.1 *The ability to train the risky arm enhances the value iff* $\kappa < \overline{\kappa}(\pi)$*, where* $\overline{\kappa}(\pi) \equiv \pi V'_{bench}(p')(1-p')$ *and* $p' \equiv \arg\max_{p\in(\hat{p},1)} V'_{bench}(p)(1-p)$.

Proof. It is optimal to train the arm for some parameter space if it is optimal to train it for at least a single belief for at least a short instance dt. Denote this unique belief as p'. At this belief, the agent must be indifferent between training and some other action, so training is weakly dominated by experimentation for any p, and $V(p) = V_{bench}(p)$, as defined in Lemma 2.

A single indifference point cannot lie inside the stopping region $(0, \hat{p})$. There, experimentation is strictly dominated by safe arm and training is strictly dominated by experimentation: $s > p\lambda + p\lambda\big(V(1) - V(p)\big) - V'(p)\lambda p(1-p) > p\lambda - \kappa + p\lambda\big(V(1) - V(p)\big) + V'(p)(\pi - \lambda p)(1-p)$, where the latter inequality relies on $V'(p) = 0$ for $p \in (0, \hat{p})$. Hence, at p' the agent is indifferent between experimenting with and without training, which holds iff $V'_{bench}(p)\pi(1-p) = \kappa$.

As $V_{bench}(p)$ is strictly convex, $V'_{bench}(p)\pi(1-p)$ is quasi-concave for $p \in [\hat{p}, 1]$, with $V'_{bench}(p)\pi(1-p) = 0$ for $p = 1$ and $p \leq \hat{p}$ (as $V'_{bench}(\hat{p}) = 0$). As such, $V'_{bench}(p')\pi(1-p') = \kappa$

identifies p' uniquely only at $\max_{p\in[\hat{p},1]} V'_{bench}(p)(1-p)$ (solution is well defined and unique due to quasi-concavity of objective function). Moreover, the above guarantees that $p' \in (\hat{p}, 1)$ - strictly bounded away from $\hat{p}$ and 1. Finally, $\overline{\kappa}(\pi) \equiv \pi V'_{bench}(p')(1-p')$ defines the upper bound on the cost for training to remain attractive: any $\kappa < \overline{\kappa}(\pi)$ implies the training region under $V'(p)\pi(1-p) \geq \kappa$ is non-empty. Note that V_{bench} and p' are independent of π, κ by construction, so $\overline{\kappa}(\pi)$ is a linear function of π. ■

Lemma B.2 (Conjecture on structure) *If $\kappa < \overline{\kappa}(\pi)$, there always exist two cutoffs $\underline{p} \in [0, p')$ and $\overline{p} \in (p', 1]$, such that the agent trains the risky arm for beliefs in $[\underline{p}, \overline{p}]$, uses the safe arm for $p < min\{\underline{p}, \hat{p}\}$ and purely experiments otherwise.*

The pure experimentation range of beliefs $(\hat{p}, \underline{p})$ is non-empty for $\kappa \in (\underline{\kappa}(\pi), \overline{\kappa}(\pi))$, where $\underline{\kappa}(\pi)$ is a non-decreasing function, and $\forall \pi, \underline{\kappa}(\pi) < \overline{\kappa}(\pi)$.

Proof.

Given the nature of conjecture, here I provide a more informal motivation for the proposed structure. I then fully prove that the conjecture is accurate in the proofs of main propositions.

$$rV(p) = \max_{(\alpha,\beta)\in[0,1]^2} \big((1-\alpha)s + \alpha(p\lambda - \beta\kappa)\big) + \alpha p\lambda\big(V(1) - V(p)\big) + V'(p)\alpha(\beta\pi - \lambda p)(1-p)$$

The optimal solution always involves a pure action ($\alpha = 0$; $\alpha = 1, \beta = 0$ or $\alpha = 1, \beta = 1$) due to linearity of $V(p)$ in α, β. Trivially, $V(p)$ is a non-decreasing function, implying that there exists a unique cutoff such that $\alpha = 0$ below this cutoff and $\alpha = 1$ above. Denote this cutoff as p^*.

By lemma B.1, under high training cost $\overline{\kappa}(\pi)$ the agent trains the arm at a unique belief $p' \in (\hat{p}, 1)$. This implies that she also trains the arm at p' for any $\kappa < \overline{\kappa}(\pi)$. By continuity of $V(p)$, for any $\kappa < \overline{\kappa}(\pi)$ there also must exist some neighbourhood around p' such that $V'(p)\pi(1-p) > \kappa$ holds there, implying that the agent is willing to train for at least a short instance dt within this neighbourhood, and never outside. Denote this region by $[\underline{p}, \overline{p}]$, with $\underline{p} < p'$ and $\overline{p} > p'$; it must be a single interval around p' by the logic above.

Note that $\overline{p}$ can go all the way up to $\overline{p} = 1$: If $\kappa = 0$, $V'(p)\pi(1-p) \geq 0$ holds even for $p = 1$, so training is weakly preferred to pure experimentation.

Similarly, under $\kappa = 0$, there always exists $\overline{\pi}$, such that $s < V'(0)\pi$ for $\pi > \overline{\pi}$ and the agent trains the arm even at $p = 0$, implying that $\underline{p} = 0$. Suppose the agent trains at $p = 0$. Then, $V(0) \geq y(0) = \frac{\pi}{r+\pi}\left(\frac{\lambda}{r} - \frac{\kappa}{r+\lambda}\right) - \frac{\kappa}{r+\pi}$ (as $C_{Tr}g(p) \geq 0$). With $\kappa = 0$, $y(0) \geq \frac{s}{r}$ iff $\pi \geq r\frac{s}{\lambda}/(1 - \frac{s}{\lambda})$ - well defined cutoff for π. Note that $p^* = \underline{p}$ in this case.

Finally, consider very high cost: $\kappa = \overline{\kappa}(\pi) - \epsilon$, where $\epsilon \to 0$. Under such cost, the training region should be very narrow, $\underline{p} \to p'_-$ and $\overline{p} \to p'_+$. This implies that $\underline{p} > \hat{p}$ holds, as $p' > \hat{p}$

by lemma B.1. Hence, there exists some cost region $(\underline{\kappa}(\pi), \overline{\kappa}(\pi))$, such that $\underline{p} > \hat{p}$, and this region is non-empty for any value of π (as least $\kappa = \overline{\kappa}(\pi) - \epsilon$ belongs to the interval). $\underline{\kappa}(\pi)$ is defined via $\underline{p} = \hat{p}$.

Note that for $\kappa \in (\underline{\kappa}(\pi), \overline{\kappa}(\pi))$, $p^* = \hat{p}$: For $p < \underline{p}$, $dp \leq 0$, so the agent's incentives match the benchmark ones according to lemma 2 in full. Then, the stopping cutoff is defined by $\hat{p}$ exactly. Overall, $p^* = \min\{\underline{p}, \hat{p}\}$, and there may exist a non-empty pure experimentation region $(\hat{p}, \underline{p})$ for high enough costs. ■

Proposition 1 (part i) and Proposition 2:

Proof. Conjecture the structure according to lemma B.2, with non-empty $(\hat{p}, \underline{p})$, and $\frac{\pi}{\lambda} < \underline{p}$. This implies that $dp \leq 0$ for any p. Then, solve for the cutoffs consecutively from the lowest to highest.

At $\hat{p}$, value matching ensures $V(\hat{p}) = \frac{s}{r}$, which is rearranged to $C_{Exp}(\hat{p}) = (\frac{s}{r} - \frac{\lambda}{r}\hat{p})\frac{1}{f(\hat{p})}$. Note that it also implies that (1) $s = \lim_{p \to \hat{p}_+}[p\lambda + p\lambda\big(V(1) - V(p)\big) - V'(p)\lambda p(1-p)]$ - a value matching in differential form, with the right limit indicating the side where the agent experiments. Marginal indifference implies that the agent does not benefit (or lose) from a short deviation to experimentation at $\hat{p}$, which holds iff (2) $s = \lim_{p \to \hat{p}_-}[p\lambda + p\lambda\big(V(1) - V(p)\big) - V'(p)\lambda p(1-p)]$ - with the left limit due to $dp < 0$ at $\hat{p}$, and $\lim_{p \to \hat{p}_-} V'(p) = 0$ as $V(p) = \frac{s}{r}$ then. Combining (1) and (2) results in $\lim_{p \to \hat{p}_-} V'(p) = \lim_{p \to \hat{p}_+} V'(p)$ - that is, smooth pasting. So $V'(\hat{p})$ is well defined, and $V'(\hat{p}) = 0$ solves for $\hat{p} = \frac{sr}{\lambda(\lambda - s + r)}$.

By similar argument, smooth pasting holds at $\underline{p}$ and $\overline{p}$. It also directly translates to $V'(p)\pi(1-p) = \kappa$ at both cutoffs. Then, the value matching pins down the constants, $C_{Tr}(\hat{p}, \underline{p}) = \big(\frac{\lambda}{r}\underline{p} - y(\underline{p}) + C_{Exp}(\hat{p})f(\underline{p})\big)\frac{1}{g(\underline{p})}$ and $C_{Exp}(\hat{p}, \underline{p}, \overline{p}) = \big(y(\overline{p}) - \frac{\lambda}{r}\overline{p} + C_{Tr}(\hat{p}, \underline{p})g(\overline{p})\big)\frac{1}{f(\overline{p})}$, and smooth pasting solves for $\underline{p}$ and $\overline{p}$. Note that $\underline{p} > \hat{p}$, as conjectured, as $V'(\underline{p})\pi(1-\underline{p}) = \kappa$ holds only if $V'(\underline{p}) > V'(\hat{p}) = 0$. Also, $\underline{p} > \frac{\pi}{\lambda}$ is necessarily satisfied, as $V(\underline{p})$ is finite, while $\lim_{p \to \frac{\pi}{\lambda}}[y(p) + C_{Tr}(\hat{p}, \underline{p})g(p)] = +\infty$, confirming the initial conjecture. Hence, the solution is characterised by:

$$V(p) = \begin{cases} \frac{s}{r} & p < \hat{p} \text{ (safe arm)} \\ \frac{\lambda}{r}p + (\frac{s}{r} - \frac{\lambda}{r}\hat{p})\frac{f(p)}{f(\hat{p})} & p \in [\hat{p}, \underline{p}] \text{ (pure exp)} \\ y(p) + \left(\frac{\lambda}{r}\underline{p} - y(\underline{p}) + (\frac{s}{r} - \frac{\lambda}{r}\hat{p})\frac{f(p)}{f(\hat{p})}\right)\frac{g(p)}{g(\underline{p})} & p \in [\underline{p}, \overline{p}] \text{ (training)} \\ \frac{\lambda}{r}p + \left(y(\overline{p}) - \frac{\lambda}{r}\overline{p} + \left(\frac{\lambda}{r}\underline{p} - y(\underline{p}) + (\frac{s}{r} - \frac{\lambda}{r}\hat{p})\frac{f(p)}{f(\hat{p})}\right)\frac{g(\overline{p})}{g(\underline{p})}\right)\frac{f(p)}{f(\overline{p})} & p > \overline{p} \text{ (pure exp)} \end{cases}$$

Finally, continuity, smoothness and convexity of $V(p)$ follow directly from value matching and smooth pasting. ■

Proposition 1 (part ii) and Proposition 3:

Proof. Conjecture the structure according to lemma B.2, and $\frac{\pi}{\lambda} > \underline{p}$. Now, $dp \leq 0$ does not hold, so break the solution down into two cases.

i. Suppose $\frac{\pi}{\lambda} < \overline{p}$. This implies that when the agent trains, her beliefs converge to $\frac{\pi}{\lambda}$, which is also inside training region, and $dp = 0$ there; i.e. the agent never switches to other options in the absence of news after she starts training and $C_{Tr} = 0$ (check that continuation value from training until news arrives and switching to pure exploitation of risky arm immediately after is captured by $y(p)$ precisely).

At $\overline{p}$, value matching and smooth pasting must hold, by the argument as in the proof of Proposition 1 (part i) and Proposition 2. Thus, $C_{Exp}(\overline{p}) = \left(y(\overline{p}) - \frac{\lambda}{r}\overline{p}\right)\frac{1}{f(\overline{p})}$ and $\overline{p} = \frac{r\left(\frac{\pi}{r+\pi}\left(\frac{\lambda}{r}-\frac{\kappa}{r+\lambda}\right)-\frac{\kappa}{r+\pi}\right)}{\kappa+r\left(\frac{\pi}{r+\pi}\left(\frac{\lambda}{r}-\frac{\kappa}{r+\lambda}\right)-\frac{\kappa}{r+\pi}\right)}$, and $\overline{p}$ falls with κ (also, for $\kappa = 0$, $\overline{p} = 1$, as conjectured). Denote $\frac{r\left(\frac{\pi}{r+\pi}\left(\frac{\lambda}{r}-\frac{\kappa}{r+\lambda}\right)-\frac{\kappa}{r+\pi}\right)}{\kappa+r\left(\frac{\pi}{r+\pi}\left(\frac{\lambda}{r}-\frac{\kappa}{r+\lambda}\right)-\frac{\kappa}{r+\pi}\right)} \equiv \overline{p}(\pi,\kappa)$ - used in later proofs.

The lower bound $\underline{p}$ is determined purely by value matching. Specifically, define the solution to $\frac{s}{r} = y(p)$ as $p^*(\pi,\kappa)$. Solving the equation gives $p^*(\pi,\kappa) = \frac{\frac{s}{r}-\left(\frac{\pi}{r+\pi}\left(\frac{\lambda}{r}-\frac{\kappa}{r+\lambda}\right)-\frac{\kappa}{r+\pi}\right)}{\frac{\lambda}{r+\pi}(1+\frac{\kappa}{r+\lambda})}$. If $p^*(\pi,\kappa) \leq \hat{p}$, this gives the solution to the lower bound, $\underline{p} = \max\{0, p^*(\pi,\kappa)\}$. $\frac{s}{r} = \left(\frac{\pi}{r+\pi}\left(\frac{\lambda}{r}-\frac{\kappa}{r+\lambda}\right)-\frac{\kappa}{r+\pi}\right) = y(0)$ defines $\kappa = \kappa_0(\pi)$, such that for $\kappa < \kappa_0(\pi)$ the following holds: $\underline{p} = 0$.

If $p^*(\pi,\kappa) > \hat{p}$, $\underline{p}$ solves for the lower root of $y(p) = V_{bench}(p)$ and it results in $\underline{p} > p^*(\pi,\kappa)$, implying that $\underline{p} > \hat{p}$. The agent purely experiments in $(\hat{p},\underline{p})$. Note that $\underline{p}$ is defined by $y(p) = V_{bench}(p)$, as the incentives of the agent for $p < \underline{p}$ coincide with the ones in the benchmark lemma 2 (she never trains there), so the value function below $\underline{p}$ coincides with $V_{bench}(p)$. The pure experimentation region $(\hat{p},\underline{p})$ is non-empty, whenever $p^*(\pi,\kappa) > \hat{p}$, so $\underline{\kappa}(\pi)$ in Lemma B.2 is defined via $p^*(\pi,\kappa) = \hat{p}$.

Note that value matching at $\underline{p}$ also guarantees that the marginal indifference holds. If the agent marginally deviates to training at $\underline{p}$, her belief will increase as $dp > 0$, and so the continuation value of such marginal deviation is $\lim_{p\to\underline{p}_+}[p\lambda + p\lambda\big(V(1)-V(p)\big) - V'(p)\lambda p(1-p)]$ - coincides with the value to the right of the cutoff. That is, the marginal indifference fully coincides with value matching, and $\lim_{p\to\underline{p}_-} V'(p) = \lim_{p\to\underline{p}_+} V'(p)$ need not hold.

In fact, smooth pasting must be violated. For any $p < \underline{p}$, training hurts, so training in dt is unattractive as well, i.e. $V'(p)\pi(1-p) \leq \kappa$ for $p < \underline{p}$. This means that (1) $\lim_{p\to\underline{p}_-}[V'(p)\pi(1-p)] \leq \kappa$. For $p \in (\underline{p},\overline{p})$, $V'(p) = const$, as the value function is linear. Hence, $V'(p)\pi(1-p)$ is a linearly decreasing function, which is equal to κ at $\overline{p}$, given smooth pasting holds there. That is, (2) $\lim_{p\to\underline{p}_+}[V'(p)\pi(1-p)] > \kappa$, as $\underline{p} < \overline{p}$. (1) and (2) together imply that $\lim_{p\to\underline{p}_-} V'(p) < \lim_{p\to\underline{p}_+} V'(p)$ - so there must be a convex kink at $\underline{p}$.

Finally, define the solution to $\underline{p}$ in this case by $\underline{p}(\pi,\kappa)$. That is,

$$\underline{p}(\pi,\kappa) \equiv \begin{cases} \max\{0, p^*(\pi,\kappa)\} & p^*(\pi,\kappa) \leq \hat{p} \\ \arg\min\{y(p) = V_{bench}(p)\} & p^*(\pi,\kappa) > \hat{p} \end{cases}$$

The conjecture of solution i. holds iff $\frac{\pi}{\lambda} \in [\underline{p}(\pi,\kappa), \underline{p}(\pi,\kappa)]$. Summarizing the above, within this parametric restriction the solution is given by:

$$V(p) = \begin{cases} \max\{\frac{s}{r}, \frac{\lambda}{r}p + (\frac{s}{r} - \frac{\lambda}{r}\hat{p})\frac{f(p)}{f(\hat{p})}\} & p < \underline{p} \text{ (safe arm or pure exp)} \\ y(p) & p \in [\underline{p}, \overline{p}] \text{ (training)} \\ \frac{\lambda}{r}p + \big(y(\overline{p}) - \frac{\lambda}{r}\overline{p}\big)\frac{f(p)}{f(\overline{p})} & p > \overline{p} \text{ (pure exp)} \end{cases}$$

and $\{\underline{p}, \overline{p}\} = \{\underline{p}(\pi,\kappa), \overline{p}(\pi,\kappa)\}$.

ii. Suppose $\frac{\pi}{\lambda} > \overline{p}$. This implies that when the agent trains, $dp > 0$ and her beliefs converge to $\overline{p}$. Similarly, being above $\overline{p}$, she purely experiments and converges to $\overline{p}$ as well. As such, $\overline{p}$ is fully absorbing, and $V(p)$ depends solely on $\overline{p}$ (and not on $\underline{p}$ or $\hat{p}$) for $p > \underline{p}$. At $\overline{p}$ the agent is indifferent between training and not, so any β is optimal. However, the unique stationary solution occurs iff $\beta^* = \overline{p}/\frac{\pi}{\lambda} \in (0,1)$ as $\frac{\pi}{\lambda} > \overline{p}$ - this guarantees that $dp = 0$ at $\overline{p}$, so the agent remains at the cutoff once converges there.

At $\overline{p}$, value matching and smooth pasting must hold. Value matching, once using differential form, implies $\lim_{p\to\overline{p}_-}[p\lambda - \kappa + p\lambda\big(V(1) - V(p)\big) + V'(p)(\pi - \lambda p)(1-p)] = \lim_{p\to\overline{p}_+}[p\lambda + p\lambda\big(V(1) - V(p)\big) - V'(p)\lambda p(1-p)]$ (limits as specified as the agent trains for $p < \overline{p}$ and purely experiments for $p > \overline{p}$). This simplifies to (1) $\lim_{p\to\overline{p}_-}[V'(p)(\pi - \lambda p)(1-p)] - \kappa = \lim_{p\to\overline{p}_+}[-V'(p)\lambda p(1-p)]$. For the marginal incentive, at $\overline{p}$ the agent is indifferent between training and not for a marginal instance dt and following the optimal strategy right after. This implies that $\lim_{p\to\overline{p}_+}[p\lambda - \kappa + p\lambda\big(V(1) - V(p)\big) + V'(p)(\pi - \lambda p)(1-p)] = \lim_{p\to\overline{p}_-}[p\lambda + p\lambda\big(V(1) - V(p)\big) - V'(p)\lambda p(1-p)]$, with limits reflecting that marginal training increases p and marginal pure experimentation decreases it. The condition simplifies to (2) $\lim_{p\to\overline{p}_+}[V'(p)(\pi - \lambda p)(1-p)] - \kappa = \lim_{p\to\overline{p}_-}[-V'(p)\lambda p(1-p)]$. Combining (1) and (2) results in $\lim_{p\to\overline{p}_-} V'(p) = \lim_{p\to\overline{p}_+} V'(p)$ - smooth pasting, so $V'(\overline{p})$ is well defined.

Value matching and smooth pasting at $\overline{p}$ allow to define C_{Tr} and C_{Exp}. Specifically, $C_{Exp}(\overline{p}) = (y(\overline{p}) - \frac{\lambda}{r}\overline{p} + C_{Tr}(\overline{p})g(\overline{p}))\frac{1}{f(\overline{p})}$ and $C_{Tr}(\overline{p}) = \left(\frac{\lambda}{r}\overline{p} - y(\overline{p}) + \frac{(\lambda\pi(1-\overline{p}) - \kappa r)\overline{p}}{r\pi(\overline{p} + \frac{r}{\lambda})}\right)\frac{1}{g(\overline{p})}$. Once plugged in, this implies that $V_{Tr}(p;\overline{p}) = y(p) + \left(\frac{\lambda}{r}\overline{p} - y(\overline{p}) + \frac{(\lambda\pi(1-\overline{p}) - \kappa r)\overline{p}}{r\pi(\overline{p} + \frac{r}{\lambda})}\right)\frac{g(p)}{g(\overline{p})}$ and $V_{Exp}(p;\overline{p}) = \frac{\lambda}{r}p + \left(\frac{(\lambda\pi(1-\overline{p}) - \kappa r)\overline{p}}{r\pi(\overline{p} + \frac{r}{\lambda})}\right)\frac{f(p)}{f(\overline{p})}$. Maximizing these with respect to $\overline{p}$ translates to $\max_{\overline{p}} C_{Exp}(\overline{p})$ or $\max_{\overline{p}} C_{Tr}(\overline{p})$ equivalently. Solving maximization problem defines the upper cutoff im-

plicitly with $\frac{\kappa}{\pi(1-\overline{p})} = \frac{\lambda}{(r+\lambda\overline{p})^2}(r+\lambda-\frac{\kappa}{\pi}r)$. Note that $\overline{p} < \frac{\pi}{\lambda}$ is necessarily satisfied, as $\lim_{p\to\frac{\pi}{\lambda}} V_{Tr}(p;\overline{p}) = +\infty$, while $V(\overline{p};\overline{p})$ is finite (and $V(p;\overline{p})$ is an increasing function). Hence, such solution satisfies the initial conjecture.

There are no improvements to the proposed solution via mixed strategies. Define $V_{Indiff}(q) \equiv V(q;q) = \frac{\lambda}{r}q + \left(\frac{(\lambda\pi(1-q)-\kappa r)q}{r\pi(q+\frac{r}{\lambda})}\right)$ - the value from mixing between experimentation with and without training (the agent mixes only if indifferent, so value matching and smooth pasting must be satisfied). $V_{Indiff}(q)$ is a concave function, while $V(p)$ is strictly convex around $\overline{p}$. $V_{Indiff}(q)$ and $V(p)$ have a single intersection at $\overline{p}$, by construction, implying that $V_{Indiff}(p) < V(p)$ for any p except $\overline{p}$ - guaranteeing no further improvements by mixing.

To conclude the solution, $\underline{p}$ is obtained solely by value matching, as in case i. I.e. $\underline{p} = \arg\min_p[V_{bench}(p) = V_{Tr}(p;\overline{p})]$. Define $\kappa_0(\pi)$ and $\underline{\kappa}(\pi)$ as costs satisfying $\underline{p} = 0$ and $\underline{p} = \hat{p}$, respectively. Note that the value function violates smooth pasting at $\underline{p}$ by the same argument as in case i.

Overall, the solution ii. is characterised by:

$$V(p) = \begin{cases} \max\{\frac{s}{r}, \frac{\lambda}{r}p + (\frac{s}{r} - \frac{\lambda}{r}\hat{p})\frac{f(p)}{f(\hat{p})}\} & p < \underline{p} \text{ (safe arm or pure exp)} \\ y(p) + \left(\frac{\lambda}{r}\overline{p} - y(\overline{p}) + \frac{(\lambda\pi(1-\overline{p})-\kappa r)\overline{p}}{r\pi(\overline{p}+\frac{r}{\lambda})}\right)\frac{g(p)}{g(\overline{p})} & p \in [\underline{p},\overline{p}] \text{ (training)} \\ \frac{\lambda}{r}p + \left(\frac{(\lambda\pi(1-\overline{p})-\kappa r)\overline{p}}{r\pi(\overline{p}+\frac{r}{\lambda})}\right)\frac{f(p)}{f(\overline{p})} & p > \overline{p} \text{ (pure exp)} \end{cases}$$

It holds whenever $\overline{p} < \frac{\pi}{\lambda}$. This is satisfied iff $\overline{p}(\pi,\kappa) < \frac{\pi}{\lambda}$ ($\overline{p}(\pi,\kappa)$ is defined in case i.). To see this, realize that at $\overline{p}(\pi,\kappa)$ the agent is indifferent by construction, so $\overline{p}(\pi,\kappa)$ lies on $V_{Indiff}(p)$, as it does on $y(p)$. Similarly, $\frac{\pi}{\lambda}$ is another intersection of $V_{Indiff}(p)$ and $y(p)$. $V_{Indiff}(\frac{\pi}{\lambda})$ is constant for any β by definition, including $\beta = 1$. At $p = \frac{\pi}{\lambda}$ and $\beta = 1$, $dp = 0$, so the agent is stuck and receives the value of training forever at full intensity until the news arrives - this is exactly $y(\frac{\pi}{\lambda})$. Given concavity of $V_{Indiff}(p)$, $\overline{p}(\pi,\kappa)$ and $\frac{\pi}{\lambda}$ are the only intersections of $V_{Indiff}(p)$ and $y(p)$, and $V_{Indiff}(p) > y(p)$ only for $p \in (\overline{p}(\pi,\kappa), \frac{\pi}{\lambda})$. At the same time, $V(\overline{p}) > y(\overline{p})$. Combining this with $V(\overline{p}) = V_{indiff}(\overline{p})$ ensures that $\overline{p} \in (\overline{p}(\pi,\kappa), \frac{\pi}{\lambda})$.

Cases i. and ii. are mutually exclusive: case i. holds if $\frac{\pi}{\lambda} \leq \overline{p}(\pi,\kappa)$, and case ii. holds if $\frac{\pi}{\lambda} > \overline{p}(\pi,\kappa)$. Finally, continuity and convexity of $V(p)$, as well as its smoothness everywhere but $\underline{p}$, follow directly in both cases from the proof provided. ■