Economics Department Discussion Papers Series

ISSN 1473 – 3307

Strategic Experimentation with Heterogeneous Agents and Payoff Externalities

Kaustav Das

Paper number 13/15

URL: http://business-school.exeter.ac.uk/economics/papers/
URL Repec page: http://ideas.repec.org/s/exe/wpaper.html

Strategic Experimentation with Heterogeneous Agents and Payoff Externalities *

Kaustav Das†

September 28, 2013

Abstract

In this paper, the problem of researchers tending to duplicate their efforts too much is addressed in models of strategic experimentation in continuous time. I consider two such models to show the conditions under which the finding of too much duplication is robust to the structure of the particular model. The first model has two arms (to be interpreted as approaches) only one of which has a chance of yielding a positive payoff to the first researcher to succeed. The second model has two independent arms; one arm is safe in the sense that it is known to finally lead to the prize; the other has a chance of getting to success much faster but it could also lead to failure. Players' abilities differ across arms. The paper falls in the literature on two-armed bandits; unlike most papers in this area we have heterogeneous players and payoff externalities.

*This is based on chapter 1 of my Ph.D dissertation written at the Pennsylvania State University. Earlier versions of the paper were circulated under the titles *Competition in R&D and Sharing of Innovative Knowledge* and *Competition, Duplication and Learning in R&D*. I thank my advisor Kalyan Chatterjee for his helpful suggestions and for providing me with continuous encouragement. I also thank Vijay Krishna, Sven Rady, Qingmin Liu, Venky Venkateswaran, David Kelsey, Dieter Balkenborg, Rajiv Sarin, James Jordan, Tymofiy Mylovanov, Neil Wallace, and Alexander Monge-Naranjo for their helpful comments. Pathikrit Basu was kind enough to go through this draft in detail and make intricate comments. Thanks are also due to the participants at the 4th World Congress of the Game Theory Society (held at Istanbul), SSCW (held at New Delhi) and Cornell-Penn State Workshop (held at Ithaca), where previous versions of this paper was presented. Finally, thanks are due to the participants of the Wednesday Theory Meet at the Penn State University, where this work has been presented at its various stages of development. The remaining errors are my own responsibility.

†Department of Economics, University of Exeter Business School, Email:K.Das@exeter.ac.uk

JEL Classification Numbers:C73, D83, O31.
Keywords: Two-armed Bandit, R&D competition, Duplication, Learning

1 Introduction

In this paper, I address the problem of optimal behavior of players in a game of strategic experimentation with two-armed bandits where there are both informational and payoff externalities as well as heterogeneous players.

In the economics literature, the two-armed bandit models have been extensively used to formally address the issue of trade-offs between exploration and exploitation in dynamic decision problems with learning. In the standard continuous time exponential bandit model, an agent has to decide how long to experiment along an arm to get rewarded before switching to experimenting along another arm. As the agent experiments along a particular arm without getting rewarded, the likelihood he attributes to ever getting rewarded along that arm is revised downwards. In this paper, I study models of strategic experimentation that incorporate variants of this standard exponential bandit with two arms. Both informational and payoff externalities are present in the models and players are heterogeneous. Informational externalities arise from the fact that an agent's learning about the state of the reward process along an arm is not only influenced by his own experimentation experiences but also by the behavior of other agents. On the other hand, payoff externalities imply that the extent to which an agent can convert a reward into a meaningful payoff depends on the order in which he gets the reward with respect to other agents. Finally, heterogeneous players mean that players have different innate abilities along different arms. Given that a reward occurs along an arm, the expected time required to get that reward differs among agents. With these features, I show that in a game of strategic experimentation, the non-cooperative equilibrium (markovian) always involves inefficient experimentation. The inefficiency is in the form of too much duplication. This means that there arise instances when all agents experiment along the same arm, though the social planner would have prefered the agents to diversify their experimentation along different arms.

In the model(s) I analyse, there are two players. Each of them faces a common continuous time two-armed bandit. Each arm can be accessed by both the players.

There can be two kinds of arms: *risky*(R) and *safe*(S). A risky arm can either be good or bad. On a good risky arm, the player who keeps on accessing it, experiences an arrival (reward occurs) almost surely. This arrival follows a Poisson process. Players differ with respect to their innate abilities which is defined by the Poisson arrival rate of breakthroughs along a good risky arm. No arrival is ever experienced along a bad risky arm. A *safe* arm is one where if a player keeps on accessing it, will almost surely experience an arrival according to a Poisson process. The intensity of the Poisson process along the safe arm is identical across players. Only the first arrival (reward) along any of the arms yields a payoff of one unit to the player who experiences it.

This paper considers two alternative settings, the purpose of which is to show that with heterogeneous players and payoff externalities, the phenomenon of too much duplication is robust to the structure of the particular model. In the first setting, both the arms are risky and are perfectly negatively correlated. Along the first good risky arm(denoted as R_1), the Poisson intensity of arrival is higher for player 1 and along the second one(denoted as R_2), if good, the intensity is higher for player 2. In the second setting there are two independent arms. One of the arms is *safe*(S) and the other is *risky*(R). Player 1's Poisson intensity of learning along the good risky arm is better than that of player 2. However, for both the players, the intensity along the good risky arm is strictly higher than that along the safe arm. At the beginning of the game, players hold a common belief about the types of the risky arm(s). Thus, in the first setting they know that with probability p, R_1 is good. Given the assumption of perfect negative correlation, this implies that they believe that with probability $(1-p)$, R_2 is good. In the second setting, players start off with a common belief about the quality of the risky arm. Each player has to decide in a continuous time regarding which arm to access. Players' actions and arrivals are publicly observable. Hence, players update their beliefs in the light of their experimentation experiences and their posterior beliefs always agree.

In the first setting, we start with analysing the social planner's problem, which aims to maximise the sum of the expected surplus of the players. The planner, in a continuous time, decides on allocating players to one of the arms. The social optimal involves *specialisation* for extreme range of beliefs and *diversification* for interim range of beliefs. This means that if it is too likely that one of the arms is good (in this setting this implies belief being close to either 0 or 1), then both the players are made

to access this arm. For interim beliefs, each player is allocated to the arm where he is better off, conditional on the arm being good. If players are homogeneous, then the range of beliefs over which players are made to access different arms, shrinks to a point which is equal to $\frac{1}{2}$.

For the analysis of noncooperative solutions, we restrict ourselves to Markovian strategies with the common posterior belief as the state variable[1]. In the present setting, the implementation of the Markovian Strategy needs special care. This is because the direction of the movement of beliefs depends on the action profile since a change in the action profile can change the direction in which the beliefs get updated. This issue is resolved by assuming that if a player is indifferent between accessing the arms, then he accesses the one where he is relatively better off. This ensures that there is always a well defined law of motion for the posterior beliefs. This assumption serves the same purpose as the idea of admissibility in Klein and Rady (2011).

The first main result shows that there cannot be an efficient equilibrium when players differ with respect to their innate abilities[2]. There always exists an equilibrium where both players use a cut-off strategy. If the extent of heterogeneity between the players(given other parameters of the model) is high enough, then the equilibrium in cutoff strategies is unique and is symmetric in nature. Otherwise, there is a multiplicity of equilibria in cut-off strategies where players change their actions at the same belief. Each of these equilibria are of self fulfilling type. However, non-cooperative equilibrium always reflects the phenomenon of too much duplication. This means that there exists a range of beliefs over which players access the same arm, when efficiency would have required one of them (the one who is better off along the other arm) to access the other arm. Interestingly, it is shown that this phenomenon of too much duplication is not just an artefact of this particular environment. It is also reflected by the results obtained in the second setting.

The second setting, as described above, is identical to the one extensively used in the existing strategic bandit literature, except for the facts that here only the first breakthrough or arrival (reward) yields a payoff and players differ with respect to their innate abilities. The planner's solution is qualitatively similar to that in the

[1]In the present model, the state of Markovian strategies should include both belief and the arm the opponent is accessing. However in the body of the paper we concentrate only on the effect of beliefs. In the supplemental appendix (available with the author) we show that this does not really matter.

[2]For homogeneous players, the equilibirum in cut-off strategies is efficient

previous setting. When it is very likely (unlikely) that the risky arm is good(bad)[3], both the players are made to access the risky (safe) arm. For interim range of beliefs, the player who is relatively better off along a risky arm is made to access it and the other player is made to access the safe arm. The noncooperative equilibrium in cut-off strategies is unique. It involves too much duplication in the sense that there exists a range of beliefs over which the less efficient player still accesses the risky arm when efficiency would require him to access the safe arm. As before, we find that for homogeneous players, the equilibrium is always efficient.

Thus, this phenomenon of too much duplication is a general characteristic of the non-cooperative equilibria of a bandit model with payoff externalities and players with different innate abilities. Inefficiency is in the form of too much experimentation along the risky arm(s), unlike in the form of free riding problem as in most of the existing strategic bandit literature.

In real world, there are many instances where agents have alternative potential approaches to pursue in order to achieve the same goal and they compete for success. Consider a situation where competing agents who are trying to make the same discovery, have a choice between potential alternate methods and the rent accruing to the second inventor is disproportionately lower than the first. This is true in many contexts. We can think of two firms engaged in a R&D race, who have alternate research methods or hypothesis to pursue. Firms do not know which method would lead to success. However, they are aware of a likelihood by which each avenue could lead to success. In this regard, one can cite an example from the pharmaceutical industry, where firms are competing to invent a drug for the *Alzheimer's* disease. Firms know that either eliminating the *beta*-amyloid protein or the *tau*-protein would eradicate the disease. Hence, firms need to decide on which hypothesis to adopt and over time they learn about the quality of the methods in the light of their search experiences. Given the high perceived valuation of a possible drug, it is evident that whoever invents the drug first would make a disproportionately higher amount of money than the later inventor(s). One could also think of a situation where two researchers are attempting to explain a scientific phenomenon. There may be alternative forms of explanation, any of which might or might not be correct. At a time there could be only one correct explanation. For example, in the $17th$ century, the Phlogiston theory used to be put forward to explain the process of combustion. However, by the end

[3]that is for high (low) beliefs

of the eighteenth century this theory was challenged and finally became void when the new Calorific theory came in. There could be similar situations in a firm also. Consider a manager who has two or more employees under his control. The manager needs to get an assignment done and would reward the employee in form of a bonus to the one who does it first. The employees have to choose among several alternate avenues to get the assignment done, although they are not sure which avenue would finally lead to success. In this case it is possible that one of the avenues will surely lead to success, but there is an alternate avenue which can either lead to success at a faster rate or can lead to failure. Clearly here each of the employees competes with others to be the first one to do the assignment successfully. In all the above situations it could be possible that conditional on an avenue being the correct one, agents would differ in their probability to achieve success along that avenue. For instance, in the pharmaceutical industry example, it is quite possible that one firm may be relatively more efficient in eradicating the β-amyloid protein, while the other may be more proficient in eradicating the τ-protein. The models of strategic experimentation analysed in this paper capture the main features of the situations described above. There are stylized facts in reality which might be due to the phenomenon of too much duplication. Again, consider the Alzheimer's drug research case. It was widely believed that the level of β-amyloid protein is the main culprit. Consequently for the past two decades almost exclusive attention was given to developing drugs to remove amyloid plaques. However, not much success has been attained in this direction. The drugs which are presently in the market, only delay the onset of this disease.([8]) As a consequence of this, the theory that β-amyloid protein is the culprit is waning and the conjecture that *tau*-proteins are to be blamed is gaining ground. However major R&D activities still involve removal of amyloid plaques. This may be due to too much duplication.

Related Literature: This paper contributes to the strategic bandit literature. Some of the works which have studied the bandit problem in the context of economics, are Bolton and Harris ([4]) Keller,Rady and Cripps([11]), Keller and Rady([12]), Klein and Rady ([14]) and Thomas([21]). In all of these papers except ([21]) and ([14]), players have replicas of bandits and *Free-riding* is a common feature in all the above models except ([21]). This leads to an inefficient level (too little) of experimentation. The present work differs from ([11]) and ([12]) in two ways. First, we have

payoff externalities. Due to this, the phenomenon of free riding does not arise . Secondly, agents differ with respect to their innate abilities. This gives us inefficiency in equilibrium, the nature of which is very different from the ones in ([11]) and ([12]).

Thomas([21]) analyses a set-up where each player has access to an exclusive risky arm, and both of them have access to a common safe arm. At a time the safe arm can be accessed by one player only. Hence, there is congestion along an arm. The present paper differs from this in the way that here each of the arms can be accessed by all the players. Further, we do not have congestion along any of the arms.

The model analysed in Klein and Rady([14]) has each player having a bandit with a safe arm and a risky arm. The risky arm of one player is perfectly negatively correlated to the risky arm of the other player. The first setting in the present paper differs from this in the following way. We have two arms, both of which are risky and perfectly negatively correlated(there is no safe arm). Each arm can be accessed by all the players. Klein([13]) addresses a model where players have replicas of bandits with three arms. One of the arms is safe and the other two are risky. The risky arms are perfectly negatively correlated. Thus, there is no payoff externalities between the players as in the present work. Also, he considers only homogeneous players.

Players in the present paper differ with respect to their innate abilities, which is absent in ([21]), ([14]) and ([13]). Evidently, this seems to be the first successful attempt in the strategic bandit literature, which explicitly works out models that incorporate difference in *learning* abilities of the players along an arm[4]. Of course we only analyse settings with payoff externalities.

This paper also contributes to the relatively less explored area of the broad literature on R&D races. It shows that in presence of heterogeneity and competition among agents, there is always a distortion in the choice of research avenue in a non-cooperative interaction. Bhattacharya and Mookerjee([3]), Dasgupta and Maskin([6]) are two of the early papers which explore this issue in a static framework. Chatterjee and Evans ([5]) analyses similar issues in a dynamic setting. The first setting of this paper has similarities with [5]. However, we consider a continuous time framework

[4]Klein and Rady([14]) discuss this issue in their work. This feature is also present in the work of Akcigit and Liu ([1]). They analyse a R&D competition model (with winner takes all structure) using two-armed bandit model with one risky and one safe arm. The risky arm could potentially lead to dead end. However their work is solely concerned about dealing with private arrival of information. I analyse this issue of private arrival of information in a different manner in another paper

with heterogeneous players. Here we can show that we always have too much duplication in the non-cooperative interaction. Some other papers to look into similar issues are Fershtman and Rubinstein ([9]) and Akcigit and Liu([1]). ([9]) studies a two-stage model in which agents simultaneously rank a finite set of boxes. Exactly one of the boxes contains the prize. Players commit to opening the boxes according to their ranked order. Inefficiency arises due to the fact that the box which is most likely to have the prize is not opened first. Their model is basically static in nature. Hence, the present paper lays down dynamic models which show that inefficiency in R&D with respect to choice of research method, is in form of too much duplication.

The rest of the paper is organised as follows. Section 2 lays down the detail of setting 1 and section 3 does the same for setting 2. Finally, section 4 concludes the paper.

2 Environment 1

2.1 The Model

There are two players 1 and 2, both of whom face a common continuous time two-armed bandit. Each of the arms is accessible by both the players. The bandit is of exponential type and both the arms are risky. Each arm can either be good or bad. On a good risky arm, a player who activates it, experiences arrivals according to a Poisson process with a known intensity. No arrivals are ever experienced on a bad risky arm. The risky arms (denoted as R_1 and R_2) are perfectly negatively correlated. Hence, one and only one of them is good. Only the first arrival (in any of the arms) yields a lump sum payoff of 1 unit to the player who experiences it.

Players differ with respect to their innate abilities. Player i $(i = 1, 2)$ experiences arrival on a good R_j $(j = 1, 2)$ according to a Poisson process with intensity $\pi_i^j > 0$. We have

$$\pi_1^1 = \pi_2^2 = \pi^{'} > \pi = \pi_1^2 = \pi_2^1$$

These Poisson intensities are common knowledge and it is evident that there is some form of symmetry among the players.

Beliefs: Players start with a given common prior p_0, which is the probability

with which R_1 is good. Their actions are publicly observable and so does any arrival experienced by them. This implies that at each date $t > 0$, players share a common posterior, denoted as p_t. If over the time interval $[t, t + \Delta)$, $\Delta > 0$, both players activate R_1 and no arrival is experienced, then using Bayes' rule, players' belief at $t + \Delta$ is

$$p_{t+\Delta} = \frac{p_t \exp^{-(\pi+\pi')\Delta}}{p_t \exp^{-(\pi+\pi')\Delta} + 1 - p_t}$$

The above expression is decreasing in Δ. The longer the players experiment on R_1 without a success, the more pessimistic they become about R_1 being good. When R_1 is activated over the time interval $dt \to 0$ (such that the terms of order $o(dt)$ can be ignored) by both the players and only unsuccessful trials are produced, the law of motion followed by the belief is then

$$dp_t = -(\pi + \pi')p_t(1 - p_t)\, dt$$

Similarly, if both the players activate R_2 over the time interval $dt \to 0$ and produce only unsuccessful trials, the law of motion followed by the belief is

$$dp_t = (\pi + \pi')p_t(1 - p_t)\, dt$$

The positive sign implies that longer the players experiment on R_2 without a success, more optimistic they become about R_1 being good. This is equivalent to become more pessimistic about R_2 being good because of the perfect negative correlation between the arms.

Given the symmetry of the model, there is no change in beliefs when each arm is activated by one player only and no arrival is experienced. For example, suppose player 1 activates R_1 and 2 activates R_2 over the time interval $[t, t + \Delta)$, without producing a success. Then, because of player 1's experimentation, the belief gets updated downwards and because of player 2's experimentation, the belief gets updated upwards. As $dt \to 0$, we can infer that the total change in belief is given by

$$dp_t = -(\pi')p_t(1 - p_t)\, dt + (\pi')p_t(1 - p_t)\, dt = 0$$

This explains why the belief gets frozen if each player activates an exclusive arm.

2.2 Social Planner's problem: The Efficient Benchmark

In this sub-section we discuss the social planner's problem which is intended to be the efficient benchmark of the model described above. In the following sub-section we would examine if non-cooperative interactions could achieve the efficient benchmark, and if not, what is the nature of the distortion.

The social planner wants to maximise the expected discounted sum of the payoffs obtained by the players. To achieve this, for each belief, he allocates a player to an arm. As stated above, only the first arrival yields a payoff. We show below that the socially optimal policy is to allocate both the players to R_1 if the belief exceeds a threshold (say p_h), both to R_2 if the belief is below a threshold (say p_l) and player 1 to R_1 and player 2 to R_2 if the belief lies in the interval $[p_l, p_h]$.

The planner's problem could be seen as a dynamic decision problem of a single player, who has at his disposal players 1 and 2, and can completely control the actions of the players. Formally, at each date $t \geq 0$, the planner chooses which option to activate from the set $\{R_1, R_2, R_1R_2\}$, where R_i $(i = 1, 2)$ means both the players are made to activate R_i and R_1R_2 involves making player 1 to activate R_1 and 2 to activate R_2. Belief p_t summarises the state and based on the trials it is updated using Bayes' rule.

At a date t, the planner's choice is summarised by the action profile $k_t = (k_{1t}, k_{2t})$. k_{it} $(i = 1, 2)$ can take values in $\{0, 1\}$ only. $k_{it} = 1$ implies that the planner makes both the players to activate R_i. $k_{1t} = k_{2t} = 0$ implies that player 1 is made to activate R_1, and 2 is made to activate R_2. Hence we can see that either $k_{1t} + k_{2t} = 1$ or $k_{1t} + k_{2t} = 0$. $k_t (t \geq 0)$ is such that it is measurable with respect to the information available at the time point t.

Assumption 1 *If the planner is indifferent between making* 1 *(*2*) to activate* R_1 *and* R_2*, then he makes* 1 *(*2*) to activate* R_1 *(*R_2*). Since in the current set-up beliefs can move in both directions, this ensures a well-defined solution to the corresponding law of motion for posterior beliefs. This is closely related to the admissibility assumption in ([14]) and ([13]).*

The expected discounted payoff to the planner can be expressed as:

$$E[\int_0^\infty e^{-rt}[(1 - k_{1t} - k_{2t})\pi' + k_{1t}p_t(\pi + \pi') + k_{2t}(1 - p_t)(\pi + \pi')]e^{X(t)}\, dt],$$

where

$$X(t) = -[\int_0^t \{(1 - k_{1\tau} - k_{2\tau})\pi^{'} + k_{1\tau}p_\tau(\pi + \pi^{'}) + k_{2\tau}(1 - p_\tau)(\pi + \pi^{'})\}\, d\tau]$$

and the expectation is over the stochastic processes k_t and p_t. Since the evolution of beliefs depends on k only, the planner's problem reduces to choosing the action profile $k = (k_1, k_2)$, given the current belief p. This implies that we can take the belief to be our state variable. Hence, we have a dynamic programming problem with the current belief p (from now on we will do away with the time subscript) as the state variable. Let $v(p)$ be the value function associated with this problem. By Bellman's principle of optimality, the value function $v(p)$ solves the following dynamic programming problem: For all $p \in [0, 1]$

$$v(p) = \max_{k_1,k_2\in\{0,1\};k_1+k_2\leq 1} \{(1 - k_1 - k_2)\pi^{'}\, dt + (k_1 p + k_2(1 - p))(\pi + \pi^{'})\, dt$$

$$+(1-r\, dt)[1-k_1 p(\pi+\pi^{'})\, dt-k_2(1-p)(\pi+\pi^{'})\, dt-(1-k_1-k_2)\pi^{'}\, dt][v(p)+v^{'}(p)(k_2-k_1)p(1-p)(\pi+\pi^{'})\, dt]\}$$

where the discount factor $\exp^{-r\, dt}$ has been approximated to $(1 - r\, dt)$ and $v(p + dp)$ and dp are substituted with $v(p)+v^{'}(p)\, dp$ and $(k_2-k_1)p(1-p)(\pi+\pi^{'})\, dt$ respectively.

After simplifying and rearranging the terms, we obtain

$$rv = \max\{B^{R_1}(v(p)), B^{R_2}(v(p)), B^{R_{12}}(v(p))\} \tag{1}$$

with

$$B^{R_1}(v(p)) := (\pi + \pi^{'})p(1 - v - (1 - p)v^{'})$$

$$B^{R_2}(v(p)) := (\pi + \pi^{'})(1 - p)(1 - v + pv^{'})$$

$$B^{R_{12}}(v(p)) := \pi^{'}(1 - v)$$

We solve the planner's problem formally in Appendix (A) and the optimal behavior is summarised in the following lemma.

Lemma 1 *For $p > p_h$, it is optimal to make both the players to activate R_1 and*

$$v(p) = \frac{\pi + \pi^{'}}{(r + \pi + \pi^{'})} p + (1-p)(\frac{1-p}{p})^{\frac{r}{\pi+\pi^{'}}} (\frac{p_h}{1-p_h})^{\frac{r+\pi+\pi^{'}}{\pi+\pi^{'}}} [\frac{\frac{\pi^{'}}{r+\pi^{'}} - \frac{\pi+\pi^{'}}{r+\pi+\pi^{'}} p_h}{p_h}], \quad (2)$$

$$\textit{where } p_h = \frac{\pi^{'}}{\pi + \pi^{'}} \textit{ and}$$

For $p < p_l$, it is optimal to make both the players to activate R_2 and

$$v(p) = \frac{\pi + \pi^{'}}{r + \pi + \pi^{'}} (1-p) + p(\frac{p}{1-p})^{\frac{r}{\pi+\pi^{'}}} (\frac{1-p_l}{p_l})^{\frac{r+\pi+\pi^{'}}{\pi+\pi^{'}}} [\frac{\frac{\pi^{'}}{r+\pi^{'}} - \frac{\pi+\pi^{'}}{r+\pi+\pi^{'}}(1-p_l)}{(1-p_l)}], \quad (3)$$

$$\textit{where } p_l = \frac{\pi}{\pi + \pi^{'}};$$

For $p \in [p_l, p_h]$, making 1 (2) to activate $R_1 (R_2)$ is optimal and

$$v(p) = \frac{\pi^{'}}{r + \pi^{'}} \quad (4)$$

The optimal associated action profile k is $(k_1, k_2) = (1, 0)$ for $p \in (p_h, 1]$; $(k_1, k_2) = (0, 1)$ for $p \in [0, p_l)$ and $(k_1, k_2) = (0, 0)$ for $p \in [p_l, p_h]$.

Consider the expression for v when $p > p_h$. The first term, $\frac{\pi+\pi^{'}}{r+\pi+\pi^{'}} p$, is the payoff from making both the firms to play R_1 for ever. However, since the arms are perfectly negatively correlated, the planner can always guarantee himself a payoff of $\frac{\pi^{'}}{r+\pi^{'}}$ by making 1 to activate R_1 and 2 to activate R_2. Hence, while making both the players to activate R_1, the planner always has the choice to switch player 2 from R_1 to R_2. This ability to switch player 2 from R_1 to R_2 is what we call the option value and it is reflected by the second term of the expression for v. It is decreasing in p, i.e. as the planner becomes more pessimistic about the quality of R_1, the payoff from having each player activating an exclusive arm increases. It can be observed that for $p = 1$, this option value is zero. This is because when it is known with certainty that R_1 is good, the payoff from making any player to activate R_2 is zero. Similarly we can explain the expression for v when $p < p_l$. This explains the solution to the planner's problem.

In the following subsection we analyse the noncooperative game and examine the nature of distortion which might arise with respect to the efficient benchmark.

2.3 The Noncooperative Game

In this subsection, we analyse the noncooperative game played by players 1 and 2. Player $i = 1, 2$ chooses actions $\{k_{i,t}\}_{t\geq 0}$, such that $k_{i,t} \in \{(1,0),(0,1)\}$ is measurable with respect to the information available at time t. $k_{i,t} = (1,0)$ indicates that player i activates R_1 and $k_{i,t} = (0,1)$ indicates that player i is activating R_2. At the beginning of the game, players hold a common prior belief about R_1 being good. This probability is exogenously given and chosen by the nature. Throughout the game, players perfectly observe each other's actions and arrivals. This implies that players share a common posterior belief at all times. As stated earlier, player getting the first arrival gets a payoff of 1 unit. We denote the players' probability assessment at time t that R_1 is good by p_t.

Throughout our analysis, we restrict players to Markovian strategies with the common belief p_t as the state variable. We define a (Markovian)strategy of player i ($i = A, B$) to be the mapping $k_i : [0,1] \to \{(1,0),(0,1)\}$ (i.e from states p_t to k_{it}. We will avoid the time subscripts from now on). We allow only those $k_i(.)$ functions which satisfy the property that $k_i^{-1}[(1,0)]$ and $k_i^{-1}[(0,1)]$ are disjoint unions of a finite number of non-degenerate sub-intervals in $[0,1]$. Also $k_i(0) = (0,1)$ and $k_i(1) = (1,0)$. These ensure that player i chooses the dominant action under subjective certainty. Given this, we can also think of the strategies of the players as follows. A player, given the current belief, chooses an arm to activate and a posterior belief at which it is going to switch and start activating the other arm. Player i's Markovian strategy is called a *threshold type* strategy if $k_i^{-1}(1,0) = [\bar{p}_1, 1]$ or $(\bar{p}_1, 1]$. Similarly player 2's Markov strategy is a *threshold type* strategy if $k_2^{-1}(0,1) = [0, \bar{p}_2]$ or $[0, \bar{p}_2)$. Finally, (k_1, k_2) is said to be symmetric if $k_1(p) = (1,0) \Leftrightarrow k_2(1-p) = (0,1)$ and $k_1(p) = (0,1) \Leftrightarrow k_2(1-p) = (1,0)$.

It should be noted that strictly speaking, the domain of a Markovian strategy of a particular player should not only depend on the current belief p, but also on the action of the other player. The supplemental appendix[5] illustrates that the results obtained by restricting the strategies of players as function of beliefs only remain valid even when strategies depend on both the belief and the action of the other player.

Assumption 2 *We assume that k_1 is right continuous and k_2 is left continuous. This guarantees the existence of a well-defined solution to the law of motion for posterior*

[5]This is available with the author

beliefs.

The above assumption means that if a player is indifferent between using R_1 or R_2, then he uses the arm along which, conditional on the arm being good, he can get arrivals with higher Poisson intensity.

Payoffs: Given an action profile (k_1, k_2), let v_1 and v_2 be the payoffs to player 1 and 2 respectively. If (k_1, k_2) constitutes an equilibrium then given k_2, v_1 along with k_1 should satisfy

$$v_1(p) = \max_{k_1 \in \{(1,0),(0,1)\}} \{(k_1^1 k_2^2 p\pi^{'}\, dt + k_1^1 k_2^1 p\pi^{'}\, dt + k_1^2 k_2^2 (1-p)\pi\, dt + k_1^2 k_2^1 (1-p)\pi\, dt)$$

$$+(1 - r\, dt)(1 - k_1^1 k_2^2 \pi^{'}\, dt - k_1^1 k_2^1 p(\pi + \pi^{'})\, dt - k_1^2 k_2^2 (1-p)(\pi + \pi^{'})\, dt - k_1^2 k_2^1 \pi\, dt)(v_1(p, k_2)$$

$$+(k_1^2 k_2^2 - k_1^1 k_2^1) p(1-p)(\pi + \pi^{'}) v_1^{'}(.)\, dt)\}$$

Expanding, rearranging and ignoring the terms of the order $o(\, dt)$, we get the following Bellman equation

$$rv_1 = \max_{k_1 \in \{(1,0),(0,1)\}} \{k_1^1 k_2^2 [\pi^{'}(p - v_1)] + k_1^1 k_2^1 [(\pi + \pi^{'}) p(\frac{\pi^{'}}{\pi + \pi^{'}} - v_1 - (1-p) v_1^{'})]$$

$$+ k_1^2 k_2^2 [(\pi + \pi^{'})(1-p)(\frac{\pi}{\pi + \pi^{'}} - v_1 + p v_1^{'})] + k_1^2 k_2^1 [\pi((1-p) - v_1)]\} \qquad (5)$$

Similarly given k_1, player 2's equilibrium payoff v_2 along with k_2 should satisfy

$$v_2(p) = \max_{k_2 \in \{(1,0),(0,1)\}} \{(k_2^2 k_1^1 (1-p)\pi^{'}\, dt + k_1^1 k_2^1 p\pi\, dt + k_1^2 k_2^2 (1-p)\pi^{'}\, dt + k_1^2 k_2^1 p\pi\, dt)$$

$$+(1 - r\, dt)(1 - k_1^1 k_2^2 \pi^{'}\, dt - k_1^1 k_2^1 p(\pi + \pi^{'})\, dt - k_1^2 k_2^2 (1-p)(\pi + \pi^{'})\, dt - k_1^2 k_2^1 \pi\, dt)(v_2(p)$$

$$-(k_1^1 k_2^1 - k_1^2 k_2^2) p(1-p)(\pi + \pi^{'}) v_2^{'}(.)\, dt)\}$$

Rearranging and simplifying (ignoring terms of the order $o(\, dt)$) the above we obtain the following Bellman equation

$$rv_2 = \max_{k_2 \in \{(1,0),(0,1)\}} \{k_2^2 k_1^1 [\pi^{'}((1-p) - v_2)] + k_2^2 k_1^2 [(\pi + \pi^{'})(1-p)(\frac{\pi^{'}}{\pi + \pi^{'}} - v_2 + p v_2^{'})]$$

$$+ k_2^1 k_1^1[(\pi+\pi^{'})p(\frac{\pi}{\pi+\pi^{'}} - v_2 - (1-p)v_2^{'})] + k_2^1 k_1^2[\pi(p-v_2)]\} \tag{6}$$

In this sub-section, the first result we present posits that the efficient benchmark can never be achieved as an outcome of a Markovian equilibrium. The following proposition states this.

Proposition 1 *There does not exist an efficient equilibrium.*

Proof. First we identify the action profile which would lead to the efficient outcome. From (1), this is as follows:

$$k_1 = (1,0) \text{ for } p \in [p_l, 1]; k_1 = (0,1) \text{ for } p \in [0, p_l)$$

$$k_2 = (1,0) \text{ for } p \in (p_h, 1]; k_2 = (0,1) \text{ for } p \in [0, p_h]$$

If this action profile would have constituted an equilibrium, the associated payoffs to player 1 and 2 would be

$$v_1(p) = \begin{cases} \frac{\pi^{'}}{r+\pi+\pi^{'}}p + C_{11}^1(1-p)[\Lambda(p)]^{\frac{r}{\pi+\pi^{'}}} & : \quad \text{if } p \in (p_h, 1], \\ \frac{\pi^{'}}{r+\pi^{'}}p & : \quad \text{if } p \in [p_l, p_h], \\ \frac{\pi}{r+\pi+\pi^{'}}(1-p) + C_{22}^1[\Gamma(p)]^{\frac{r}{\pi+\pi^{'}}} & : \quad \text{if } p \in [0, p_l) \\ & : \end{cases} \tag{7}$$

and

$$v_2(p) = \begin{cases} \frac{\pi}{r+\pi+\pi^{'}}p + C_{11}^2(1-p)[\Lambda(p)]^{\frac{r}{\pi+\pi^{'}}} & : \quad \text{if } p \in (p_h, 1], \\ \frac{\pi^{'}}{r+\pi^{'}}(1-p) & : \quad \text{if } p \in [p_l, p_h], \\ \frac{\pi^{'}}{r+\pi+\pi^{'}}(1-p) + C_{22}^2 p[\Gamma(p)]^{\frac{r}{\pi+\pi^{'}}} & : \quad \text{if } p \in [0, p_l) \\ & : \end{cases} \tag{8}$$

where C_{11}^i , C_{22}^i $(i = 1, 2)$ are integration constants and are derived by imposing the *value matching condition* at p_h and p_l.

For the above action profile to constitute an equilibrium, k_1 should be a best response to k_2 for all $p \in [0, 1]$ and vice-versa. This implies that v_1 and v_2 (given by 7 and 8) along with the efficient action profile (k_1, k_2) should satisfy (5) and (6) respectively. Appendix (B) shows that there exists $p^{'} > p_l$, such that for $p \in (p_l, p^{'}]$, k_1 does not constitute a best response to k_2.

This concludes the proof. ■

Symmetric Inefficient equilibrium:

The efficient action profile is symmetric(in the manner described above) and involves a range of beliefs over which each player activates an exclusive arm. Since we have shown that there does not exist an efficient equilibrium, it is of natural interest to look for outcomes that can be obtained in a symmetric Markovian equilibrium of the non-cooperative game and which involves a range of beliefs over which each player activates an exclusive arm. It turns out that for certain parametric conditions, we can obtain a unique equilibrium outcome (in threshold type Markovian strategies). The following proposition describes this.

Proposition 2 *If $r(\pi^{'} - \pi) - \pi\pi^{'} > 0$, then the unique Markovian equilibrium in threshold type strategies is symmetric and is constituted by the strategy profile (k_1^N, k_2^N) given by*

$$k_1^N = (1,0) \text{ for } p \in [p_l^*, 1]; k_1^N = (0,1) \text{ for } p \in [0, p_l^*)$$

$$k_2^N = (1,0) \text{ for } p \in (p_h^*, 1]; k_2^N = (0,1) \text{ for } p \in [0, p_h^*]$$

where

$$p_l^* = \frac{\pi(r+\pi^{'})}{r\pi^{'} + \pi(r+\pi^{'})} \text{ and } p_h^* = \frac{r\pi^{'}}{r\pi^{'} + \pi(r+\pi^{'})} = 1 - p_l^*$$

$$\text{Also } p_l < p_l^* < p_h^* < p_h$$

Proof. We prove this proposition with the help of the following lemmas.

Lemma 2 *If player 2 is activating R_2 for $p \in [0, \bar{p}]$ and R_1 for $p \in (\bar{p}, 1]$ (where $\frac{1}{2} \leq \bar{p} \leq p_h$), then there exists a p_l^* satisfying $0 < p_l^* < \frac{1}{2} \leq \bar{p}$, such that for player 1, activating R_2 for $p \in [0, p_l^*)$ and R_1 for $p \in [p_l^*, 1]$ constitutes a best response to player 2's strategy.*

Proof of Lemma. By hypothesis, we have $k_2 = (0,1)$ for $p \in [0, \bar{p}]$ and $k_2 = (1,0)$ for $p \in (\bar{p}, 1]$, such that $p_h \geq \bar{p} \geq \frac{1}{2}$. We know that given this, for $p = 0$, it is optimal for 1 to choose $k_1 = (0,1)$(that is to activate R_2). We now need to find the point where 1 will find it optimal to switch to activate R_1 rather than R_2, given k_2. Hence we need to solve for the *optimal stopping problem* of player 1 in the region $[0, \bar{p}]$. Appendix (C) solves the optimal stopping problem of player 1 and shows that

$$k_1^N = (1,0) \text{ for } p \in [p_l^*, 1]; k_1^N = (0,1) \text{ for } p \in [0, p_l^*)$$

constitutes a best response to k_2, for all $\bar{p} \geq \frac{1}{2}$. ■

Next, we obtain similar kind of best response function for player 2, given player 1's strategy.

Lemma 3 *If player* 1 *is activating* R_1 *for* $p \in [\underline{p}, 1]$ *and* R_2 *for* $p \in [0, \underline{p})$ *(where* $\frac{1}{2} \geq \underline{p} \geq p_l$*), then there exists a* p_h^* *satisfying* $1 > p_h^* > \frac{1}{2}$*, such that for player* 2*, activating* R_2 *for* $p \in [0, p_h^*]$ *and* R_1 *for* $p \in (p_h^*, 1]$ *constitutes a best response to player* 1*'s strategy.*

Proof of Lemma. We have $k_1 = (1, 0)$ for $p \in [\underline{p}, 1]$ and $k_1 = (0, 1)$ for $p \in [0, \underline{p})$, such that $\underline{p} \leq \frac{1}{2}$. Given this, at $p = 1$, 2 finds it optimal to choose $k_2 = (1, 0)$ (that is to activate R_1). As before, we intend to find the point where 2 will switch to activate R_2. (the optimal stopping problem of player 2 in the region $[\underline{p}, 1]$).

Appendix (D) solves this optimal stopping problem and shows that

$$k_2^N = (1, 0) \text{ for } p \in (p_h^*, 1]; k_2^N = (0, 1) \text{ for } p \in [0, p_h^*]$$

constitutes a best response to k_1 for all $\underline{p} \leq \frac{1}{2}$. ■

The supplemental appendix obtains the value of the integration constants by imposing the value matching condition at p_h^* and p_l^*. Also it shows that v_1 and v_2 satisfy the smooth pasting condition at p_h^* and p_l^* respectively, conditional on the other player's strategy.

Since $p_h > p_h^* > \frac{1}{2}$ and $\frac{1}{2} > p_l^* > p_l$, the proof of the proposition now follows directly from lemma (2) and (3). We can posit that

$$k_1^N = (1, 0) \text{ for } p \in [p_l^*, 1]; k_1^N = (0, 1) \text{ for } p \in [0, p_l^*)$$

constitutes a best response to

$$k_2^N = (1, 0) \text{ for } p \in (p_h^*, 1]; k_2^N = (0, 1) \text{ for } p \in [0, p_h^*]$$

and vice-versa. ■

Proposition (2) characterises the non-cooperative equilibrium when $r(\pi^{'} - \pi) - \pi^{'}\pi > 0$. It is to be observed that $p_l^* > p_l$ and $p_h^* < p_h$. The inefficiency of the non-cooperative equilibrium entails from the fact that in the intervals $[p_l, p_l^*)$ and $(p_h^*, p_h]$, players activate the same arm when efficiency requires one of them to switch

to activate the other arm. Hence there exist ranges of beliefs, such that if the state lies in one such range, there is *too much* duplication with respect to experimentation along one of the arms.

Given π and r and $\pi^{'} > \pi$, the condition $r(\pi^{'} - \pi) - \pi^{'}\pi > 0$ puts a lower bound on the value of $\pi^{'}$. Hence to have unique symmetric equilibrium, we should have the degree of heterogeneity between the agents high enough. This condition can be intuitively explained as follows. Suppose there exists a range of beliefs such that player 1 activates R_1 and player 2 R_2. Over this range of diversification (i.e each player activating an exclusive arm), the payoffs to 1 and 2 are $\frac{\pi^{'}}{r+\pi^{'}}p$ and $\frac{\pi^{'}}{r+\pi^{'}}(1-p)$ respectively. If player 1 unilaterally deviates and switch to activate R_2, then the belief will be updated upwards conditional on no arrival and if 2 unilaterally deviates and switch to activate R_1, the belief will be updated downwards, conditional on no arrival. This implies that if player 1 unilaterally deviates and switch to activate R_2 over the time interval dt, then conditional on no arrival, the expected discounted future payoff will be higher for player 1. Similarly, if player 2, deviates and switch to activate R_1 over the dt time interval then conditional on no arrival its expected discounted future payoff increases.

Thus activating an exclusive arm (as described above) is incentive compatible from players' point of view, if at each p, given the other player's action, the instantaneous payoff to a player from diversification is no less than that from activating the same arm as his competitor. This is a necessity. Consider a p in such a range. Given that player 2 is activating R_2, player 1 knows that the expected discounted payoff from diversification is $\frac{\pi^{'}}{r+\pi^{'}}p$. The instantaneous payoff is $\frac{\pi^{'}}{r+\pi^{'}}pr\,dt$. The instantaneous payoff from activating the same arm as his competitor is $\pi(1-p)\,dt$. Thus it is optimal for firm 1 to activate R_1 if,

$$\frac{\pi^{'}}{r+\pi^{'}}pr\,dt \geq \pi(1-p)\,dt \tag{9}$$

Similarly, given player 1 is activating R_1, player 2 finds it optimal to activate R_2 if

$$\frac{\pi^{'}}{r+\pi^{'}}(1-p)r\,dt \geq \pi p\,dt \tag{10}$$

Thus to have a range of beliefs in a non-cooperative equilibrium over which players will activate an exclusive arm, (9) and (10) should hold together, with strict inequality

for at least one p. This implies

$$\frac{\pi^{'}}{r+\pi^{'}}r > \pi$$

$$\Rightarrow r(\pi^{'}-\pi)-\pi^{'}\pi > 0$$

This explains the condition required for the existence of a symmetric equilibrium. Thus to have an equilibrium with players activating an exclusive arm for a range of beliefs, it is necessary that the extent of arm specific heterogeneity is high enough. In this paper this is reflected by the magnitude of the term $(\pi^{'}-\pi)$. The condition is more likely to be true, when the value of $(\pi^{'}-\pi)$ is higher(for a given value of r). However one can see that for low values of r, (i.e when agents become more patient) the condition is less likely to hold.

Asymmetric Inefficient equilibria: It is clear from the previous proposition that we cannot have symmetric equilibrium when the condition $r(\pi^{'}-\pi)-\pi^{'}\pi > 0$ fails to hold. In these situations we can expect to obtain equilibria where the equilibrium strategy profile $(k_1^{N'}, k_2^{N'})$ involves switching of both players at the same belief. This implies that in these equilibria, players activate an exclusive arm only at the belief p^*, the common switching point. This follows from assumption (2).

To begin with, we focus on the benchmark case first, i.e the situation when players are homogeneous.

$$\pi_1^1 = \pi_2^1 = \pi_1^2 = \pi_2^2 = \pi$$

Clearly, the condition $r(\pi^{'}-\pi)-\pi^{'}\pi > 0$ fails to hold. The following proposition describes the equilibrium.

Proposition 3 *If $\pi^{'} = \pi$, the unique equilibrium in threshold type strategies is constituted by the strategy profile (k_1^{Ne}, k_2^{Ne}) such that $k_1^{Ne} = (1,0)$ for $p \in [p^*, 1]$ and $k_1^{Ne} = (0,1)$ for $p \in [0, p^*)$; $k_2^{Ne} = (1,0)$ for $p \in (p^*, 1]$ and $k_2^{Ne} = (0,1)$ for $p \in [0, p^*]$, where $p^* = \frac{1}{2}$.*

Proof. Consider the above proposed strategy profile. Given player 2's strategy, player 1 finds it optimal to switch to activate R_1 rather than R_2 at the belief p^*. This requires $p^* \geq \frac{1}{2}$. The proof is available in the supplemental appendix

Similarly, at p^*, given player 1's strategy, player 2 finds it optimal to switch to activate R_2 rather than R_1. This requires $p^* \leq \frac{1}{2}$. and the proof can be found in the supplemental appendix.

This implies that if there is an equilibrium with the same switching point for both the players, then the switching point should be $p^* = \frac{1}{2}$. The supplemental appendix establishes that the above proposed strategy profile with $p^* = \frac{1}{2}$ constitutes best response to each other for all $p \in [0, 1]$.

This concludes the proof. ■

The above analysis shows that in the absence of any heterogeneity among the firms, we have a unique equilibrium in threshold type Markovian strategies with players activating an unique arm at one point only. By recalling the analysis of the social planner's problem we can posit that when players are homogeneous, the outcome of this equilibrium coincides with the efficient outcome.

We now turn our focus to the situation when players are heterogeneous and we cannot have equilibrium that involves players activating an exclusive arm over a range of beliefs. The following proposition describes this.

Proposition 4 *If $\pi^{'} > \pi > 0$ and the condition $r(\pi^{'} - \pi) - \pi^{'}\pi > 0$ fails to hold, then we have a multiplicity of equilibria in threshold type strategies as described below:*

$$k_1^s = (0, 1) \text{ for } p \in [0, p^*) \text{ and } k_1^s = (1, 0) \text{ for } p \in [p^*, 1]$$

$$k_2^s = (0, 1) \text{ for } p \in [0, p^*] \text{ and } k_2^s = (1, 0) \text{ for } p \in (p^*, 1]$$

where,

$$p^* \in [\max\{\underline{p}_s, p_h^*\}, 1 - \max\{\underline{p}_s, p_h^*\}]$$

and

$$\underline{p}_s = \frac{\pi(r + \pi^{'})}{\pi(r + \pi^{'}) + \pi^{'}(r + \pi)}$$

Proof. Since the condition $r(\pi^{'} - \pi) - \pi^{'}\pi > 0$ fails to hold, we cannot have equilibrium with diversification (i.e a range of beliefs over which players activate an exclusive arm.). Thus we seek to find equilibria where the switching points for the players are the same. Let p^* be the common switching point.

At the belief p^*, player 1 finds it optimal to switch to activate R_1 from R_2. This requires (shown formally in the supplemental appendix)

$$p^* \geq \frac{\pi(r + \pi^{'})}{\pi(r + \pi^{'}) + \pi^{'}(r + \pi)} = \underline{p}_s$$

Similarly we can show that if player 2 finds it optimal to switch at p^*, then we must have

$$p^* \leq \frac{\pi^{'}(r+\pi)}{\pi(r+\pi^{'})+\pi^{'}(r+\pi)} = \bar{p}_s$$

Thus to have an equilibrium with same switching points, it is necessary that the switching point p^* lies in the interval $[\underline{p}_s, \bar{p}_s]$. Since $r(\pi^{'}-\pi)-\pi^{'}\pi \leq 0$, $p_h^* < \frac{1}{2}$. In an equilibrium where the switching points are the same it is a necessity that the switching point $p^* \geq p_h^*$. Otherwise from our previous analysis we know that if $p^* < p_h^*$ and 1 switches at p^* then 2 finds it optimal to switch at p_h^* and not p^*. Hence $p^* \in [\max\{\underline{p}_s, p_h^*\}, 1-\max\{\underline{p}_s, p_h^*\}]$.

Finally, we need to establish that k_1^s (k_2^s) constitutes a best response to k_2^s (k_1^s) for $p \in (p^*, 1]$ ($[0, p^*)$). This is established in the supplemental appendix.

This concludes the proof of the proposition ■

It is to be noted that in the present case, smooth pasting condition will not necessarily be satisfied by v_i at p^*. This is because here we are in some sense getting a *corner solution* for the optimal stopping problems of player 1 and player 2. That is, given $k_2^s = (0,1)$ for $p \in [0, p^*]$, player 1 would have ideally liked to switch to R_1 from R_2 at $p = p_l^*$. However he will not be able to do this since $p^* \leq p_l^*$ and player 2 would switch to R_1 at $p = p^*$. Similar thing will be true for player 2 as well.

The previous proposition states that when $r(\pi^{'}-\pi)-\pi^{'}\pi < 0$ and firms are heterogeneous, we have a multiplicity of equilibria where players have a common switching point. Each of this equilibria is of *self-fulfilling* type. Since p^*, the common switching point always lies in the interval (p_l, p_h)[6], each of this equilibria involves *too much* duplication. The analysis of the above model shows that whenever the players are heterogeneous,(i.e their Poisson intensities of learning along an arm differ)the non-cooperative equilibrium is inefficient, such that for a certain range of beliefs, there is too much experimentation along one of the arms when efficiency would require one of the players to conduct experimentation along the other arm. Hence the phenomenon of too-much duplication in the present set-up can be perceived as a manifestation of payoff externalities and heterogeneity among the players. In the supplemental appendix, it is shown that the results obtained in propositions (2) and (4) hold even if strategies depend both on beliefs and the action of the opponent.

Next, we analyse a different model in Environment 2 and show that too much

[6]this follows straigghtaway from the expression of $\underline{p}_s$

duplication can also generalise to other bandit models. The setting is similar to the ones analysed in ([11]) and ([12]). However here only the first arrival yields a payoff and we allow for heterogeneity among the players. Thus apart from showing that the phenomenon of too much duplication is robust to the structure of the particular model, we also show the nature of inefficiency in a bandit model with one safe arm and one risky arm in presence of payoff externalities and difference in innate abilities across the players.

3 Environment 2

Two players face a common continuous time two-armed bandit. Each of the arms is accessible by both the players. The bandit is of exponential type. One of the arms is *safe*(S) and the other arm is *risky*(R). A player who activates the safe arm, gets arrival according to a Poisson process with intensity $\pi_0 > 0$. A risky arm can either be good or bad. If player i activates a *good* risky arm, then he experiences arrival according to a Poisson process with intensity π_i, such that

$$\pi_1 \geq \pi_2 > \pi_0 > 0$$

No arrivals are experienced along a bad risky arm.

Players start with a common prior p^0, which is the probability with which the risky arm is good. Players observe each other's actions and the arrivals experienced by them. Hence at each time point t, players share a common posterior belief p_t. Only the first player to experience an arrival gets a payoff of 1 unit. We start with the case when the players are homogeneous, i.e $\pi_1 = \pi_2$.

3.1 Symmetric Players

In this subsection, we lay out the analysis with homogeneous players. Thus players' ability to learn along the risky arm is the same. They both experience arrivals at the good risky arm according to a Poisson process with intensity $\pi_1 > \pi_0$.

We start our analysis with the benchmark case, i.e the social planner's problem.

3.1.1 Social Planner's problem: The efficient benchmark

Consider the problem of a benevolent social planner who wants to maximise the sum of expected discounted payoff of the players. Hence at each instant, based on p, he allocates each of the players to activate one of the arms. k_t denotes the action profile chosen by the planner at the instant t. $k_t \in \{0, 1, 2\}$. k_t denotes the number of players made to activate the risky arm at the instant t. $k_t (t \geq 0)$ is such that it is measurable with respect to the information available at time t

It is assumed that if the planner is indifferent between making a player to activate the risky arm or the safe arm, then he makes him to activate the safe arm. Thus the planner's action is left continuous.

From now on we will do away with the time subscript. Let $v(p)$ be the value function of the planner. Since actions are left continuous and beliefs can move only in the left direction, left continuity of $v(p)$ can always be assumed.

Then $v(p)$ should satisfy,

$$v(p) = \max_{k \in \{0,1,2\}} \{(2-k)\pi_0 \, dt + kp\pi_1 \, dt +$$

$$(1 - r \, dt)(1 - (2-k)\pi_0 \, dt - kp\pi_1 \, dt)(v(p) - v^{'}(.)kp(1-p) \, dt)\},$$

since $(v(p + \, dp) = v(p) + v^{'}(p) \, dp)$ and $\, dp = kp(1-p) \, dt$.

After expanding and rearranging the above and ignoring the terms of order $o(\, dt)$ we have

$$rv = \max_{k \in \{0,1,2\}} \{(2-k)\pi_0[1-v] + k(\pi_1 p[1 - v - v^{'}(1-p)])\} \tag{11}$$

Proposition 5 *The planner's optimality involves making both the players to activate the risky arm as long as $p > p^*$, where $p^* = \frac{\pi_0}{\pi_1}$. For $p \leq p^*$, both are made to activate the safe arm.*

Proof. Since (11) is linear in k, we know that at the optimum, k will either be 2 or 0. When both players are optimally made to activate the risky arm, the value function satisfies:

$$v = \frac{2\pi_1}{r + 2\pi_1} + C(1-p)[\Lambda(p)]^{\frac{r}{2\pi_1}},$$

where $\Lambda(p) = \frac{1-p}{p}$ and C is the integration constant. This is derived by solving the O.D.E obtained by putting $k = 2$ in (11).

When both players are optimally made to activate the safe arm, then $v = \frac{2\pi_0}{r+2\pi_0}$. Since $v(p)$ satisfies the *value matching* and *smooth pasting* conditions at $p = p^*$, we get

$$C = \frac{\frac{2\pi_0}{r+2\pi_0} - \frac{2\pi_1}{r+2\pi_1}}{(1-p^*)[\Lambda(p)]^{\frac{r}{2\pi_1}}} \text{ and } p^* = \frac{\pi_0}{\pi_1}$$

This concludes the proof. ■

3.1.2 The non-cooperative game

Player i chooses actions $\{k_{it} \in \{0,1\}\}$, such that k_{it} is measurable with respect to the information available at time t. We restrict our attention to Markovian strategies, such that strategy of player i is defined by the mapping $k_i : [0,1] \to \{0,1\}$. We allow only those k_i functions which satisfy the property that $k_i^{-1}(1)$ and $k_i^{-1}(0)$ are disjoint unions of a finite number of non-degenerate sub-intervals in $[0,1]$, such that $k_i(0) = 0$ and $k_i(1) = 1$. This ensures that the game is well-defined in the continuous time framework.

Players simultaneously update their belief about the risky arm to be good as long as there is at least one player activating the risky arm and there is no arrival(at any of the arms). Both k_1 and k_2 are left continuous, which guarantee the existence of a well defined law of motion of the posterior.

Let v_i be the value function (equilibrium payoff) of player i $(i = 1, 2)$in the non-cooperative game. If (k_1, k_2) is an equilibrium strategy profile then given k_j $(j = 1, 2)$, k_i $(i = 1, 2; i \neq j)$ and v_i should satisfy

$$v_i = \max_{k_i \in \{0,1\}} \{(1-k_i)\pi_0\, dt + k_i\pi_1 p\, dt +$$

$$(1 - r\, dt)(1 - \pi_0\, dt(2 - k_i - k_j) - p\pi_i(k_i + k_j)\, dt)(v_i - v_i' p(1-p)(k_i + k_j)\, dt)\}$$

Simplifying the above, we obtain

$$rv_i = \max_{k_i \in \{0,1\}} \{(1-k_i)\pi_0(1-v_i) + k_i(\pi_1 p[1 - v_i - v_i' p(1-p)])$$

$$- (1-k_j)\pi_0 v_i - k_j\pi_j p(v_i + (1-p)v_i')\} \tag{12}$$

Proposition 6 *There exists an efficient equilibrium.*

Proof. Consider the following strategy profile: *Each player activates R for $p > p^*$ and S for $p \leq p^*$* (Hence p^* is the switching point). This is a symmetric strategy profile and the outcome implied by this profile is the efficient outcome. We need to show that this profile constitutes an equilibrium.

Suppose player 2 follows the above strategy. We will determine the best response of player 1. It is clear that for $p = 1$, player 1 will choose R. Thus the optimal switching point of player 1 is to be determined. It is shown in the supplemental appendix that the unique optimal switching point for player 1 is p^*. Similarly, this can be shown for player 2.

This concludes the proof. ■

This is an interesting point to note. From [11] we know that with homogeneous players, efficient equilibrium in threshold type strategies never exists. Here we observe that just by introducing payoff externalities, we can obtain efficient equilibrium in threshold type strategies. Hence we see that competition among players brings in efficiency which intuitively makes sense.

Next, we move on to our analysis with heterogeneous players. We find that the nature of distortion in the non-cooperative game with respect to the benchmark case (social planner's problem) is exactly the same as obtained in the previous environment.

3.2 Heterogeneous Players

Consider the setting where players are heterogeneous, i.e their ability to learn across the risky arm is different. Hence we have $\pi_1 > \pi_2 > \pi_0$.

To start with, as before, we first analyse the social planner's problem which is intended to be the efficient benchmark.

3.2.1 The Social Planner's problem

The planner's objective is the same as before. Let (k_1, k_2) be his action profile. $k_i \in \{0, 1\}$, for $i = 1, 2$. $k_i = 1(0)$ implies that the planner has made the ith player to activate risky(safe) arm. Let $v(p)$ be the value function of the planner. Then it should satisfy

$$v(p) = \max_{k_i \in \{0,1\}} \{(2 - k_1 - k_2)\pi_o \, dt + k_1 p \pi_1 \, dt + k_2 p \pi_2 \, dt +$$

$$(1-r\,dt)(1-(2-k_1-k_2)\pi_0\,dt-k_1p\pi_1\,dt-k_2p\pi_2\,dt)(v(p)-v^{'}(p)p(1-p)(k_1\pi_1+k_2\pi_2)\,dt)\}$$

$$\Rightarrow rv = \max_{k_i\in\{0,1\}}\{(2-k_1-k_2)\pi_0[1-v]+k_1(p\pi_1[1-v-v^{'}(1-p)])+k_2(p\pi_2[1-v-v^{'}(1-p)])\} \tag{13}$$

This is because $v(p+\ dp) = v(p) + v^{'}(p)\,dp$ and $\ dp = -(k_1\pi_1 + k_2\pi_2)\,dt$.

The following lemma establishes a property for an interior solution of the planner's problem.

Lemma 4 *If there exists an interior solution (i.e there exists $p_i^* \in (0,1)$ such that for $p > p_i^*$ player i is made to activate R and for $p \leq p_i^*$, player i is made to activate S) then optimality requires diversification over a range of beliefs. That is, there exists a range of beliefs over which the planner will make one player to activate the risky arm and the other player to activate the safe arm.*

Proof of Lemma. Suppose not. This implies that the planner's optimality requires him to switch both the player from the risky arm to the safe arm at the same p, say $p^{'}$. At the optimum the smooth pasting condition must hold which implies that $v^{'}(p^{'}) = 0$. From (13), we know that optimality requires,

$$p^{'}\pi_2[1-v] = p^{'}\pi_1[1-v(p^{'})] = \pi_0[1-v(p^{'})]$$

However since $\pi_1 > \pi_2$, $p^{'}\pi_2[1-v(p)] < p^{'}\pi_1[1-v(p^{'})]$. This is a contradiction.

This proves the lemma. ■

The next lemma shows that if the planner's solution involves diversification, then player 2 is to be switched to the safe arm at a higher belief than the one at which player 1 is switched.

Lemma 5 *Player 2 is to be switched to the safe arm from the risky arm at a higher p than player 1.*

Proof of Lemma. Suppose not. From lemma (4) we know that this implies player 1 is switched to the safe arm at a higher p than player 2. Let this switching point be p_1^*. From (13), we know that at p_1^* we must have, $\pi_0[1-v(p_1^*)] = p_1^*\pi_1[1-v(p_1^*)-v^{'}(p_1^*)(1-p_1^*)]$. Since $\pi_2 < \pi_1$, we have $\pi_0[1-v(p_1^*)] = p_1^*\pi_1[1-v(p_1^*)-v^{'}(p_1^*)(1-p_1^*)] > p_1^*\pi_2[1-v(p_1^*)-v^{'}(p_1^*)(1-p_1^*)]$. This is a contradiction to the claim that it is optimal to keep player 2 at the risky arm at $p = p_1^*$. This proves the lemma. ■

With the help of the above two lemmas we are now in a position to describe the planner's solution. The following proposition does this.

Proposition 7 *There exists a solution to the planner's problem, where both the players are made to activate the risky arm for $p > p_2^*$, player 2 is made to activate the safe arm and 1 to activate the risky arm for $p \in (p_1^*, p_2^*]$, and both players are made to activate the safe arm for $p \leq p_1^*$ where $p_1^* = \frac{\pi_0}{\pi_1}$.*

Proof. First, assume that there exists some $\frac{\pi_0}{\pi_1} < p_2^* < 1$, such that it is optimal to switch player 2 to the safe arm at p_2^*. $v(p)$ in the range of beliefs over which 2 is made to activate the safe arm and 1 is made to activate the risky arm, should satisfy

$$v = \frac{\pi_0}{r+\pi_0} + \frac{r\pi_1 p}{(r+\pi_0)(r+\pi_0+\pi_1)} + C_2(1-p)[\Lambda(p)]^{\frac{r+\pi_0}{\pi_1}} \equiv v_{SR}$$

This is derived through solving the O.D.E obtained by putting $k_2 = 0$ and $k_1 = 1$ in (13). Suppose p_1^* is the belief where 1 is to be switched to the safe arm. Since at p_1^*, both players are activating S, optimality would require to have $v^{'}(p_1^*) = 0$(smooth pasting condition). According to lemma (5), player 2 is switched from R to S at a higher p. Then from the value matching condition, we know that we should have $v_{SR}(p_1^*) = v(p_1^*)$. This gives us $C_2 = \frac{\frac{r\pi_0}{(r+\pi_0)(r+2\pi_0)} - \frac{r\pi_1 p_1^*}{(r+\pi_0)(r+\pi_0+\pi_1)}}{(1-p_1^*)[\Lambda(p_1^*)]^{\frac{r+\pi_0}{\pi_1}}}$. Observe that $C_2 > 0$. Also, the smooth pasting condition at p_1^* implies $v_{SR}^{'}(p_1^*) = 0$. This gives us

$$\frac{r\pi_1}{(r+\pi_0)(r+\pi_0+\pi_1)} - C_2[\Lambda(p_1^*)]^{\frac{r+\pi_0}{\pi_1}}[1 + \frac{(r+\pi_0)}{\pi_1 p_1^*}] = 0 \Rightarrow p_1^* = \frac{\pi_0}{\pi_1}$$

We now need to prove the existence of a $p_2^* \in (p_1^*, 1)$, such that at p_2^*, the planner finds it optimal to switch player 2 from R to S. The existence of such a p_2^* is proved in the supplemental appendix.

This concludes the proof of the proposition. ■

Corollary 1 *$p_2^* > \frac{\pi_0}{\pi_2}$, the threshold p where the planner would have switched player 2 from R to S had he been dealing with this player only.*

Proof. Suppose not. Then $p_2^* \leq \frac{\pi_0}{\pi_2}$. At p_2^*, $v^{'}(p_2^*) = v_{SR}^{'}(p_2^*) > 0$. Since v is strictly convex for $p > \frac{\pi_0}{\pi_1}$, $v^{'}(\frac{\pi_0}{\pi_2}) > 0$. Therefore at $p = \frac{\pi_0}{\pi_2}$, $\pi_0[1-v] > \pi_2 p[1-v-v^{'}(1-p)]$. From (13), we can see that this contradicts the claim that $p_2^* \leq \frac{\pi_0}{\pi_2}$. This proves the corollary. ■

3.2.2 The non-cooperative game

This is similar to the non-cooperative game with homogeneous players. Thus $k_1(.)$ and $k_2(.)$ are the Markovian strategies of the players.

Let $v_1(p)$ and $v_2(p)$ be the payoff functions of players 1 and 2 respectively in a Markovian equilibrium. v_i along with k_i should then satisfy

$$rv_i = \max_{k_i \in \{0,1\}} \{(1-k_i)[\pi_0(1-v_i)]+k_i[\pi_i p(1-v_i-v_i^{'}(1-p))]-[(1-k_j)\pi_0 v_i+k_j p(v_i+v^{'}(1-p))]\} \tag{14}$$

This implies that given k_j, at any p optimality on player i's part requires choosing $k_i(p) = 0(1)$ if $[\pi_0(1-v_i)] \geq (<)[\pi_i p(1-v_i-v_i^{'}(1-p))]$.

We determine the non-cooperative equilibrium in following steps.

Lemma 6 *Suppose player 2 follows the strategy of activating R for $p > p_2^{*N}$ and S for $p \leq p_2^{*N}$ such that $\frac{\pi_0}{\pi_1} < p_2^{*N} < 1$. Then player 1's best response is to activate R for $p > p_1^*$ and S for $p \leq p_1^{*N}$ where $p_1^* = \frac{\pi_0}{\pi_1}$.*

Proof of Lemma. First, consider the range $p \leq p_2^{*N}$. If $k_1 = 1$ ($k_2 = 0$ by hypothesis), then by putting $i = 1$ in (14) we know that v_1 should solve

$$v_1^{'} + \frac{[r+\pi_0+\pi_1]}{p(1-p)\pi_1} v_1 = \frac{1}{(1-p)}$$

This is a first order O.D.E. Solving this we have,

$$v_1 = \frac{\pi_1}{r+\pi_0+\pi_1} p + C(1-p)[\Lambda(p)]^{\frac{r+\pi_0}{\pi_1}} \equiv v_1^{RS}(p) \tag{15}$$

where C is an integration constant. If he choose $k_1 = 0$ then $v_1(p)$ should satisfy,

$$v_1 = \frac{\pi_0}{r+2\pi_0} \tag{16}$$

Initially, we assume that player 1 indeed behaves in the way as claimed, for $p \leq p_2^{*N}$. Later, we will show that the value function thus obtained for the specified range will satisfy the Bellman equation for this range. This is shown in the supplemental appendix.

Next, consider the range $p > p_2^{*N}$. As before we conjecture that it is optimal for 1 to choose $k_1 = 1$ and derive the value function. Then we show that the obtained value

function indeed satisfy the bellman equation. Again, this is shown in the supplemental appendix.

This concludes the proof. ■

Lemma 7 *Suppose player* 1 *plays the following strategy: Activate* R *for* $p > p_1^{*N} = \frac{\pi_0}{\pi_1}$ *and Activate* S *for* $p \leq p_1^{*N}$. *Then there exists a* $p_2^{*N} \in (p_1^{*N}, \frac{\pi_0}{\pi_2})$, *such that player* 2*'s best response is to activate* R *for* $p > p_2^{*N}$ *and activate* S *for* $p \leq p_2^{*N}$.

Proof of Lemma. Consider $p \leq p_1^{*N}$. First, as before we conjecture that it is optimal for player 2 to be at S. Then $v_2 = \frac{\pi_0}{r+2\pi_0}$ for $p \leq p_1^{*N}$. From (14) one can conclude that $\pi_0(1 - v_2) > \pi_2 p[1 - v_2 - v_2'(1-p)]$ for $p \leq p_1^{*N}$. This supports our conjecture.

Now consider the optimal stopping problem of player 2 in the range $[p_1^{*N}, 1]$, given player 1's strategy. This is done in the supplemental appendix, which shows the existence of a unique $p_2^{*N} \in (p_1^{*N}, 1)$.

From (14), we know that at the optimal we shall have $[\pi_2 p_2^{*N}(1 - v_2(p_2^{*N}) - v_2'(p_2^{*N})(1 - p_2^{*N}))] = \pi_0(1 - v_2(p_2^{*N}))$. Since $[1 - v_2(p_2^{*N})] < [1 - v_2(p_2^{*N}) - v_2'(p_2^{*N})(1 - p_2^{*N})]$, we have $p_2^{*N} < \frac{\pi_0}{\pi_2}$. ■

The above two lemmas now allow us to formally state the non-cooperative equilibrium. The following proposition describes this.

Proposition 8 *Player* 1 *activating* R *(*S*) for* $p > (\leq) p_1^{*N}$ *and player* 2 *activating* R *(*S*)for* $p > (\leq) p_2^{*N}$ *constitutes a unique Markovian equilibrium in threshold type strategies.*

Proof. The proof of this proposition follows directly from lemma (6) and (7). ■

The above proposition describes the unique equilibrium in threshold type Markovian strategies. Since $p_2^{*N} < \frac{\pi_0}{\pi_2} < p_2^*$, there exists a range of beliefs (p_2^{*N}, p_2^*) when efficiency requires player 2 to switch to the safe arm, but it does not. This shows, that the non-cooperative equilibrium outcome involves the phenomenon of too-much duplication.

Proposition (8) strengthens the notion, that in a two-armed bandit model with heterogeneous players and payoff externalities among them, non-cooperative interactions between the players involve distortions in the form of too much experimentation along one of the arms.

4 Conclusion

We have demonstrated that when the players are heterogeneous with respect to learn across the risky arm(s), then efficiency requires diversification, i.e each player to experiment along an exclusive arm, specifically the arm along which he is relatively better off in learning . This has been established in two different environments. In presence of heterogeneity among players, we do not achieve efficiency in a non-cooperative interaction. Only when the players are homogeneous, is the non-cooperative outcome efficient. When the players are heterogeneous, we always have too much duplication in a non-cooperative interaction. Depending on the parameter values we can either have a unique equilibrium with diversification over a range of beliefs or a multiplicity of equilibria with diversification at a point only.

Future research can take the following two pathways. First, it would be interesting to explore the situation when there can be private arrival of information. This means that some arrival on a risky arm is only observable to the player who experiences it. This is already being addressed in a present research by myself. Secondly, one can address the bandit problems where each arm can be independently good or bad. It would be interesting to see the nature of non-cooperative interactions in such an environment.

References

[1] Akcigit, U., Liu, Q., 2011: "The Role of Information in Competitive Experimentation. ", *mimeo, Columbia University and University of Pennsylvania.*

[2] d'Aspremeont, C., Bhattacharya, S., Gerard-Varet, L., 2000 "Bargaining and Sharing Innovative knowledge. ", *The Review of Economic Studies* 67, $255-271$.

[3] Bhattacharya, S., Mookerjee D., 1986 "Portfolio choice in research and development. ", *Rand Journal of Economics* 17, $594-605$.

[4] Bolton, P., Harris, C., 1999 "Strategic Experimentation. ", *Econometrica* 67, $349-374$.

[5] Chatterjee, K., Evans, R., 2004: "Rivals' Search for Buried Treasure: Competition and Duplication in R&D. ", *Rand Journal of Economics* 35, $160-183$.

[6] Dasgupta, P., Maskin, E., 1987: "The Simple Economics of Research Portfolios ", *The Economic Journal* 581 − 595

[7] Dasgupta, P., Stiglitz, J., 1980 "Uncertainty, Industrial Structure and the Speed of R&D ", *Bell Journal of Economics* 111 − 28

[8] "No end to Dementia ", *The Economist, June* 2010

[9] Fershtman, C., Rubinstein, A., 1997 "A Simple Model of Equilibrium in Search Procedures. ",*Journal of Economic Theory* 72, 432 − 441.

[10] Graham, M.B.W., 1986 "The Business of research ", *New York:Cambridge University Press.*

[11] Keller, G., Rady, S., Cripps, M., 2005: "Strategic Experimentation with Exponential Bandits ", *Econometrica* 73, 39 − 68.

[12] Keller, G., Rady, S., 2010:"Strategic Experimentation with Poisson Bandits ", *Theoretical Economics* 5, 275 − 311.

[13] Klein, N., 2011: "Strategic Learning in Teams ", *mimeo University of Bonn*

[14] Klein, N., Rady, S., 2011: "Negatively Correlated Bandits ", *The Review of Economic Studies* 78 693 − 792.

[15] Lee, T., Wilde, L., 1980: "Market Structure and Innovation: A Reformulation", *Quarterly Journal of Economics* 94 429 − 436

[16] Loury,G.C., 1979 "Market Structure and Innovation ", *Quarterly Journal of Economics* 93395 − 410.

[17] Presman, E.L., 1990: "Poisson Version of the Two-Armed Bandit Problem with Discounting, *Theory of Probability and its Applications*

[18] Reinganum, J. 1982 "A dynamic Game of R&D Patent Protection and Competitive Behavior ", *Econometrica* 50 671 − 688.

[19] Scherer, F.M., "International High-Technology Competition ", *Cambridge, Mass:, Harvard University Press*

[20] Stokey,N.L., 2009: "The Economics of Inaction ", *Princeton University Press.*

[21] Thomas, C., 2011: "Experimentation with Congestion ", *mimeo, University College of London and University of Texas Austin*

APPENDIX

A Proof of Lemma (1)

We prove this lemma in two steps. First, we derive the planner's optimal value function (v) assuming it to be of the threshold type. This means that the planner changes its actions at countable number of beliefs in the interval $[0, 1]$. Then, we show that the obtained v along with the associated action profile k satisfy the optimality equation given by (1).

Step 1: Deriving the $v(p)$

Suppose $p = 1$,i.e it is known with certainty that R_1 is good. Then both the players would have been made to activate R_1. Since beliefs continuously change in the leftward direction when both activate R_1, v can be assumed to be left continuous at the $\epsilon-$neighborhood of 1. Hence in the $\epsilon-$neighborhood of 1, the planner would still make both the players to activate R_1. We need to find the belief at which the planner would find it optimal to make player 2 to activate R_2 rather than R_1. Hence we need to solve for the optimal stopping problem of the planner. When both the firms are made to activate R_1, we have

$$rv(p) = B^{R_1}(v(p)) = (\pi + \pi^{'})p(1 - v - (1 - p)v^{'})$$

Solving the above, we find that v in the neighborhood of 1 satisfies the following O.D.E

$$v^{'}(p) + \frac{[r + (\pi + \pi^{'})p]}{(\pi + \pi^{'})p(1 - p)}v(p) = \frac{1}{1 - p}$$

The integrating factor $\mu(p)$ of the above differential equation is given by

$$\mu(p) = e^{\int \frac{r+(\pi+\pi^{'})p}{(\pi+\pi^{'})p(1-p)}\,dp} = \frac{p^{\frac{r}{\pi+\pi^{'}}}}{(1-p)^{\frac{r}{\pi+\pi^{'}}+1}}$$

Multiplying both sides of the O.D.E with $\mu(p)$ and integrating both sides we get

$$\int d[v(p)\frac{p^{\frac{r}{\pi+\pi^{'}}}}{(1-p)^{\frac{r}{\pi+\pi^{'}}+1}}] = \int \frac{p^{\frac{r}{\pi+\pi^{'}}}}{(1-p)^{\frac{r}{\pi+\pi^{'}}+2}} + C_1$$

$$\Rightarrow v = \frac{\pi+\pi^{'}}{r+\pi+\pi^{'}}p + C_1(1-p)[\Lambda(p)]^{\frac{r}{\pi+\pi^{'}}}$$

where C_1 is the integration constant and $\Lambda(p) = \frac{1-p}{p}$.

This is the payoff function of the planner when he makes both player 1 and 2 to activate R_1. The second term in the R.H.S of the above expression reflects the planner's choice of being able to make player 2 to activate R_2 rather than R_1. We determine the optimal belief for the planner to do this by imposing the *value matching condition* and the *smooth pasting condition* on v at that optimal belief. This way, we obtain the optimal belief $p_h = \frac{\pi^{'}}{\pi+\pi^{'}}$, at and below which the planner finds it optimal to make player 2 to activate R_2. The integral constant C_1 is also determined and substituting these, we obtain 2.

Similarly, suppose $p = 0$, i.e it is known with certainty that R_2 is good. Then both the players would have been made to activate R_2. As explained above, we need to find a belief such that when the state reaches there from the left side, the planner finds it optimal to make player 1 to activate R_1 rather than R_2. As before, the value function of the planner can be assumed to be right continuous at the $\epsilon-$neighborhood of 0. To determine the optimal belief where the planner changes his action, we again solve for the optimal stopping problem of the planner. When both the players are optimally made to activate R_2, then we have

$$rv(p) = B^{R_2}(v(p)) \equiv (\pi+\pi^{'})(1-p)(1-v+pv^{'})$$

which gives us the following first order O.D.E which v satisfies in the right neighborhood of 0

$$v^{'} - \frac{[r+(1-p)(\pi+\pi^{'})]}{p(1-p)(\pi+\pi^{'})}v = -\frac{1}{p}$$

Solving the above differential equation we obtain

$$v = \frac{\pi+\pi^{'}}{r+\pi+\pi^{'}}(1-p) + C_2(p)[\Lambda(p)]^{-\frac{r}{\pi+\pi^{'}}}$$

where C_2 is the integration constant. This gives us the payoff function of the planner when he makes both the players to activate R_2. Imposing the smooth pasting and the value matching condition, we obtain the optimal belief $p_l = \frac{\pi}{\pi+\pi'}$, at and above which the planner finds it optimal to make player 1 to activate R_1 rather than R_2. After obtaining the value of C_2 and substituting we get (3). Since $p_l < p_h$, for the beliefs in the range $[p_l, p_h]$, the planner makes player 1 to activate R_1 and player 2 to activate R_2. Then we have

$$v(p) = B^{R_1R_2} \equiv \pi'(1 - v(p))$$

This gives us (4). The action profile k associated with the obtained value function is $(k_1, k_2) = (1, 0)$ for $p \in (p_h, 1]$; $(k_1, k_2) = (0, 1)$ for $p \in [0, p_l)$ and $(k_1, k_2) = (0, 0)$ for $p \in [p_l, p_h]$.

Step 2:

Next, we show that the obtained value function in the previous step along with the associated action profile k satisfy the Bellman equation given by (1).

To begin with, consider the region $(p_1^*, 1]$. v' is given by $\frac{\pi+\pi'}{r+\pi+\pi'} - C_1(\Lambda(p))^{\frac{r}{\pi+\pi'}}[1 + \frac{r}{\pi+\pi'}.\frac{1}{p}]$ This implies

$$(1-p)v' = \frac{\pi+\pi'}{r+\pi+\pi'} - [p\frac{\pi+\pi'}{r+\pi+\pi'} + C_1(1-p)[\Lambda(p)]^{\frac{r}{\pi+\pi'}}] - C_1\frac{(1-p)}{p}[\Lambda(p)]^{\frac{r}{\pi+\pi'}}\frac{r}{\pi+\pi'}$$

$$\Rightarrow (1 - v - (1-p)v') = \frac{r}{\pi+\pi'+r} + \frac{(1-p)}{p}C_1[\Lambda(p)]^{\frac{r}{\pi+\pi'}}\frac{r}{\pi+\pi'}$$

Substituting the above in the expression of $B_1(p, v)$, we get $B_1(p, v) = rv$.

Further, from the expression of $v'(p)$ we obtain

$$pv' = \frac{\pi+\pi'}{r+\pi+\pi'}p - pC_1[\Lambda(p)]^{\frac{r}{\pi+\pi'}} - C_1\frac{r}{\pi+\pi'}[\Lambda(p)]^{\frac{r}{\pi+\pi'}} = v - C_1[\Lambda(p)]^{\frac{r}{\pi+\pi'}}.\frac{r+\pi+\pi'}{\pi+\pi'}$$

$$\Rightarrow 1 - v + pv' = 1 - C_1[\Lambda(p)]^{\frac{r}{\pi+\pi'}}.\frac{r+\pi+\pi'}{\pi+\pi'}$$

Substituting this in the expression of $B_2(p, v)$, we get $B_2(p, v) = (\pi+\pi') - (r+\pi+\pi')v$. Thus to have $B_1(p, v) \geq B(p, v)$ and $B_1(p, v) \geq B_2(p, v)$, we require $v \geq \frac{\pi'}{r+\pi'}$ and $v \geq \frac{\pi+\pi'}{2r+\pi+\pi'}$ respectively. Since $\frac{\pi'}{r+\pi'} - \frac{\pi+\pi'}{2r+\pi+\pi'} = \frac{\delta(\pi'-\pi)}{(r+\pi')(2r+\pi+\pi')} > 0$ and for $p \in (p_1^*, 1]$

$v > \frac{\pi^{'}}{r+\pi^{'}}$, we have

$$B_1(p,v) = \max\{B(p,v), B_1(p,v), B_2(p,v)\}$$

Similarly we can show that for the region $p \in [0, p_2^*)$, $B_2(p,v) = \max\{B(p,v), B_1(p,v), B_2(p,v)\}$.

Lastly, consider the region $[p_l, p_h]$. $v = \frac{\pi^{'}}{r+\pi^{'}}$. Hence $v^{'} = 0$. This gives us $B^{R_1}(.) = (\pi + \pi^{'})p[1-v]$; $B^{R_2}(.) = (\pi + \pi^{'})(1-p)[1-v]$;$B^{R_1R_2} = \pi^{'}[1-v]$. Thus

$$B^{R_1R_2} \geq B^{R_1} \Rightarrow p \leq \frac{\pi^{'}}{\pi + \pi^{'}} = p_h \text{ and } B^{R_1R_2} \geq B^{R_2} \Rightarrow p \geq \frac{\pi}{\pi + \pi^{'}} = p_l$$

Thus for $p \in [p_l, p_h]$, we have $B^{R_1R_2} = \max\{B^{R_1}, B^{R_1R_2}, B^{R_2}\}$

This shows that the value function and the corresponding policy k satisfies the Bellman equation given by (1).

B Proof of the proposition (1)

Consider the region $[p_l, p_h]$. Payoffs induced by the efficient action profile (k_1, k_2) implies that for $p \in [p_l, p_h]$, $v_1 = \frac{\pi^{'}}{r+\pi^{'}}p$. From (5), we know that to have k_1 to be a best response to k_2, we require

$$\pi^{'}(p - v_1) \geq (\pi + \pi^{'})(1-p)(\frac{\pi}{\pi + \pi^{'}} - v_1 + pv_1^{'})$$

$$\Rightarrow p \geq \frac{\pi(r + \pi^{'})}{r\pi^{'} + \pi(r + \pi^{'})} = p^{'}(\text{ say})$$

However,

$$\frac{\pi(r + \pi^{'})}{r\pi^{'} + \pi(r + \pi^{'})} - p_l = \frac{\pi(r + \pi^{'})}{r\pi^{'} + \pi(r + \pi^{'})} - \frac{\pi}{\pi + \pi^{'}}$$

$$= \frac{\pi(\pi^{'2})}{[r\pi^{'} + \pi(r + \pi^{'})][(\pi + \pi^{'})]} > 0 \Rightarrow p^{'} > p_l$$

Hence for $p \in [p_l, p^{'})$, $\pi^{'}(p - v_1) < (\pi + \pi^{'})(1-p)(\frac{\pi}{\pi+\pi^{'}} - v_1 + v_1^{'}p)$. Thus k_1 does not constitute a best response to k_2 for values of p in this region.

C Solving for player 1's best response function in lemma (2)

Let p_l^* be the switching point for player 1, i.e, the belief where he switches to activate R_1 rather than R_2. First, we assume that $p_l^* < \frac{1}{2}$. This will then induce a payoff function for 1 which satisfies (7) with p_l replaced by p_l*. Since v_1 thus obtained is a continuous function, at the switching point p_l^* we shall have,

$$\pi^{'}\{p_l^* - v_1\} = (\pi + \pi^{'})(1 - p_l^*)\{\frac{\pi}{\pi + \pi^{'}} - v_1 + p_l^* v_1^{'}\}$$

Given k_2, p can change in one direction only (which is right in this case). This implies that we can take $v_1^{'} = \frac{\pi^{'}}{r+\pi^{'}}$. Hence

$$\pi^{'}\frac{r}{r + \pi^{'}}p_l^* = (1 - p_l^*)\pi$$

$$\Rightarrow p_l^* = \frac{\pi(r + \pi^{'})}{r\pi^{'} + \pi(r + \pi^{'})}$$

Since $r(\pi^{'} - \pi) - \pi\pi^{'} > 0$, $p_l^* < \frac{1}{2}$. This is consistent with the assumption we made before. Also it can be shown that $p_l^* > p_l$. This shows that $k_1^N = (1, 0)$ is an optimal response to k_2 for $p \in [p_l^*, \bar{p}]$.

Next, Consider the region $(\bar{p}, 1]$. For $k_1^N = (1, 0)$ to constitute a best response to k_2 we must have,

$$(\pi + \pi^{'})p(\frac{\pi^{'}}{\pi + \pi^{'}} - v_1 - v_1^{'}(1 - p)) \geq \pi((1 - p) - v_1)$$

In this region, $v_1^{'}$ is given by

$$v_1^{'} = \frac{\pi^{'}}{r + \pi + \pi^{'}} - C_{11}^1(\Lambda(p))^{\frac{r}{\pi+\pi^{'}}}[1 + \frac{r}{\pi + \pi^{'}}\frac{1}{p}] \tag{17}$$

This implies,

$$(\pi + \pi^{'})p(\frac{\pi^{'}}{\pi + \pi^{'}} - v_1 - v_1^{'}(1 - p)) = rv_1$$

Thus we require

$$rv_1 \geq \pi((1 - p) - v_1) \Rightarrow v_1 \geq \frac{\pi}{r + \pi}(1 - p)$$

From the value matching condition we know that $v_1(\bar{p}) = \frac{\pi^{'}}{r+\pi^{'}}\bar{p}$. Since $\frac{1}{2} \leq \bar{p} < p_h$, from the *switching derivative lemma* (refer to appendix E) we know that $v_1^{'} > 0$ for all $p \in [\bar{p}, 1]$. Hence we must have $v_1 > \frac{\pi}{r+\pi}(1-p)$ for all $p \in (\bar{p}, 1]$. This implies that $k_1^N = (1, 0)$ is an optimal response to k_2 for $p \in (\bar{p}, 1]$.

Thus we have shown that

$$k_1^N = (1,0) \text{ for } p \in [p_l^*, 1]; k_1^N = (0,1) \text{ for } p \in [0, p_l^*)$$

constitutes a best response to k_2 for all $p \in [0, 1]$.

D Solving for player 2's best response function in lemma (3)

Let p_h^* be the switching point of player 2. Assuming $p_h^* > \frac{1}{2}$, this will induce a payoff function for 2 which satisfies (8) with p_h replaced by p_h^*. At $p = p_h^*$ we shall have

$$\pi^{'}((1-p_h^*) - v_2) = (\pi + \pi^{'})p_h^*(\frac{\pi}{\pi + \pi^{'}} - v_2 - v_2^{'}(1-p))$$

Since given k_1, p can change in one direction only (which is left in this case). Thus we can take $v_2^{'} = -\frac{\pi^{'}}{r+\pi^{'}}$. This implies

$$\pi^{'}(1-p_h^*)\frac{r}{r+\pi^{'}} = p\pi$$

$$\Rightarrow p_h^* = \frac{r\pi^{'}}{r\pi^{'} + \pi(r+\pi^{'})}$$

Since $r(\pi^{'} - \pi) - \pi\pi^{'} > 0$, $p_h^* > \frac{1}{2}$. This is consistent with our assumption made before. Also, it can be shown that $p_h^* < p_h$. Applying similar logic as shown in the previous lemma to the region $[0, \underline{p}]$, we can posit that

$$k_2^N = (1,0) \text{ for } p \in (p_h^*, 1]; k_2^N = (0,1) \text{ for } p \in [0, p_h^*]$$

constitutes a best response to k_1 for all $p \in [0, 1]$.

E Switching-derivative lemma

Lemma 8 *When both players are activating R_1 then $v_1^{'}(.)$ is given by*

$$v_1^{'}(.) = \frac{\pi^{'}}{r+\pi+\pi^{'}} - C_{11}^{1}[\Lambda(p)]^{\frac{r}{\pi+\pi^{'}}}(1+\frac{r}{\pi+\pi^{'}}\frac{1}{p})$$

Similarly when both players are activating R_2 then $v_2^{'}(.)$ is given by

$$v_2^{'}(.) = -\frac{\pi^{'}}{r+\pi+\pi^{'}} + C_{22}^{2}[\Gamma(p)]^{\frac{r}{\pi+\pi^{'}}}(1+\frac{r}{\pi+\pi^{'}}\frac{1}{1-p})$$

Let p_s^2 (p_s^1) be the switching point of 2 (1). Then if $p_s^2 < p_h$ ($p_s^1 > p_l$), $v_1^{'}(p_s^2) > 0$ ($v_2^{'}(p_s^1) < 0$).

Proof of Lemma.

Since C_{11}^{A} is chosen by imposing value matching at p_s^2, we have

$$v_1^{'}(p_s^2) = \frac{\pi^{'}}{r+\pi+\pi^{'}} - \{[\frac{\pi^{'}}{r+\pi^{'}} - \frac{\pi^{'}}{r+\pi+\pi^{'}}](\frac{p_s^2}{1-p_s^2} + \frac{r}{\pi+\pi^{'}}\frac{1}{1-p_s^2})\}$$

It is easy to see that $v_1^{'}(p_s^2)$ is decreasing in p_s^2. Also we can show that $v_1^{'}(p_h) = 0$ where $p_h = \frac{\pi^{'}}{\pi+\pi^{'}}$. Hence if $p_s^2 < p_h$ then $v_1^{'}(p_s^2) > 0$. Similarly we can argue for $v_2^{'}(p_s^1)$.

■