



Max Planck Institute  
for Innovation and Competition

# Exploration and Exploitation

Dr. Rainer Widmann

# Agenda

---

- Part 1: Models of Experimentation (in the context of project choice)
  - Pandora's Box Problem
  - Multi-Armed Bandit Model
- Part 2: Motivating exploration in organizations
  - Motivating a worker to innovate (Manso 2011)
  - Short-Termism



# Models of Experimentation

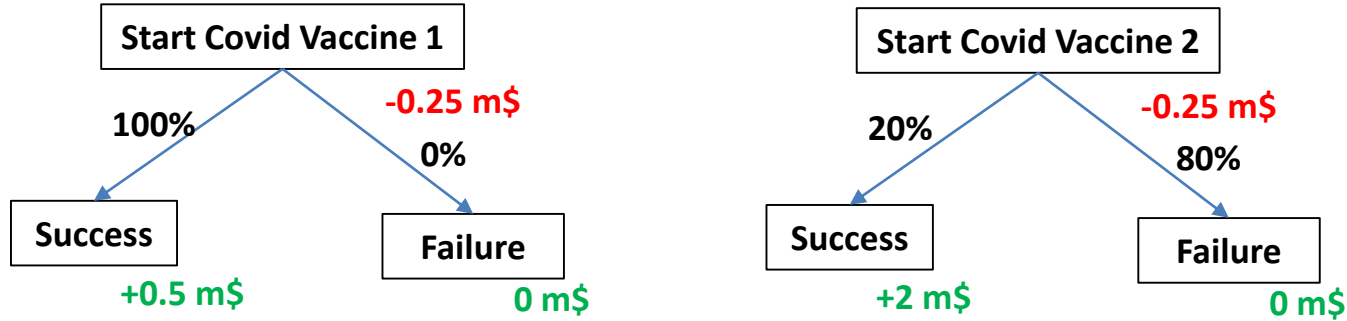
---

## The “Pandora’s box” problem

- Weitzman (1981)
  - Suppose there are  $N$  “Boxes”:  $b_1, \dots, b_N$
  - Each box has a cost of opening it:  $c_1, \dots, c_N$
  - Each box has a stochastic distribution of rewards:  $F_1, \dots, F_N$  (indep. distr.)
- Rules:
  - You may open as many boxes as you like
  - You get to keep the reward of only one box



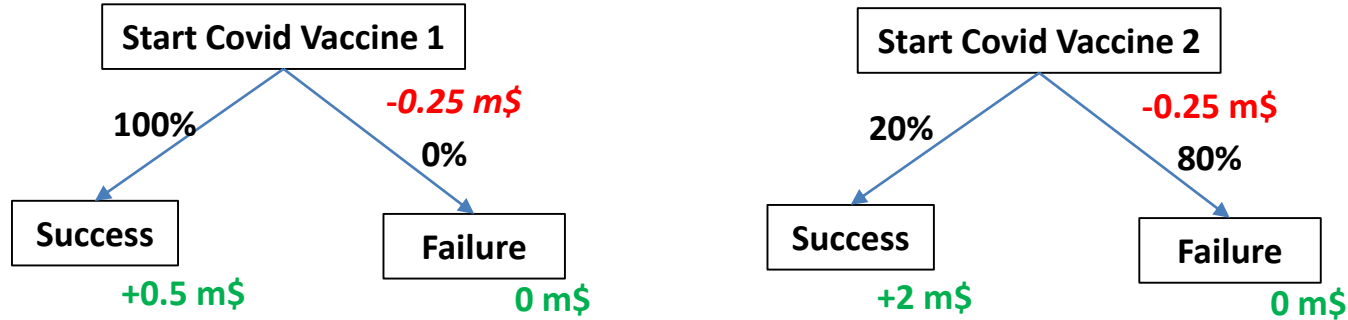
# Models of Experimentation



- **Problem 1:** What if you have to commit to undertaking one project only?



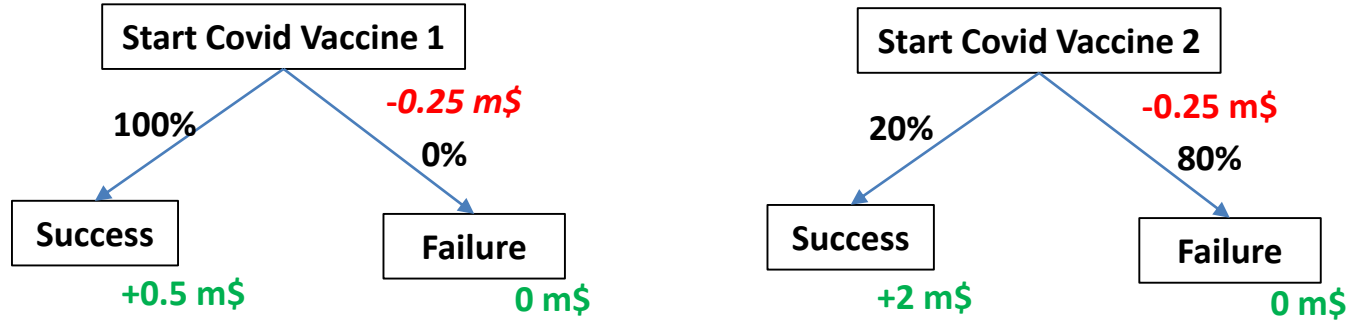
# Models of Experimentation



- **Problem 1:** What if you have to commit to undertaking one project only?
  - Solution: calculate expected value of each project (Vac1:  $0.5 \times 1 - 0.25 = \mathbf{0.25}$ , Vac2:  $2 \times 0.2 - 0.25 = \mathbf{0.15}$ ) and choose **Vac1** (simple!)
- **Problem 2:** But what if you can undertake both projects sequentially?



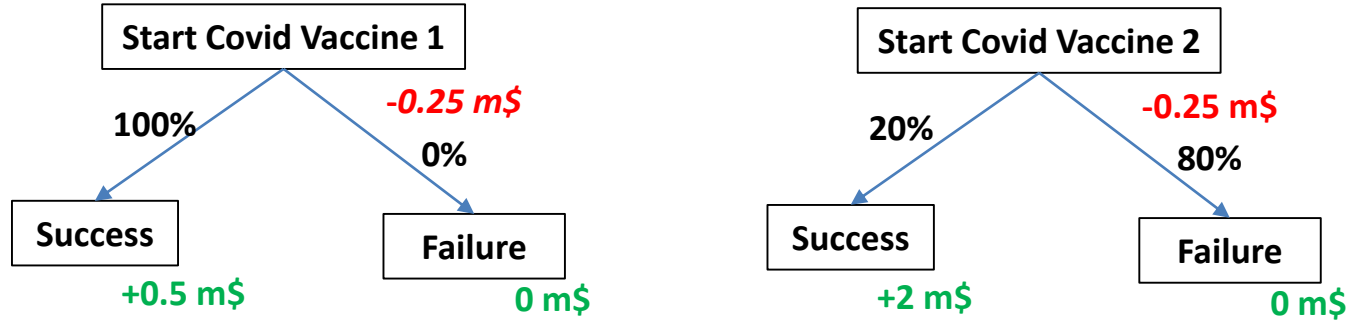
# Models of Experimentation



- **Problem 1:** What if you have to commit to undertaking one project only?
  - Solution: calculate expected value of each project (Vac1:  $0.5 \cdot 1 - 0.25 = \mathbf{0.25}$ , Vac2:  $2 \cdot 0.2 - 0.25 = \mathbf{0.15}$ ) and choose **Vac1** (simple!)
- **Problem 2:** But what if you can undertake both projects sequentially?
  - Strategy 1: Undertake Vac1, then undertake Vac2: Exp. payoff=**0.3**



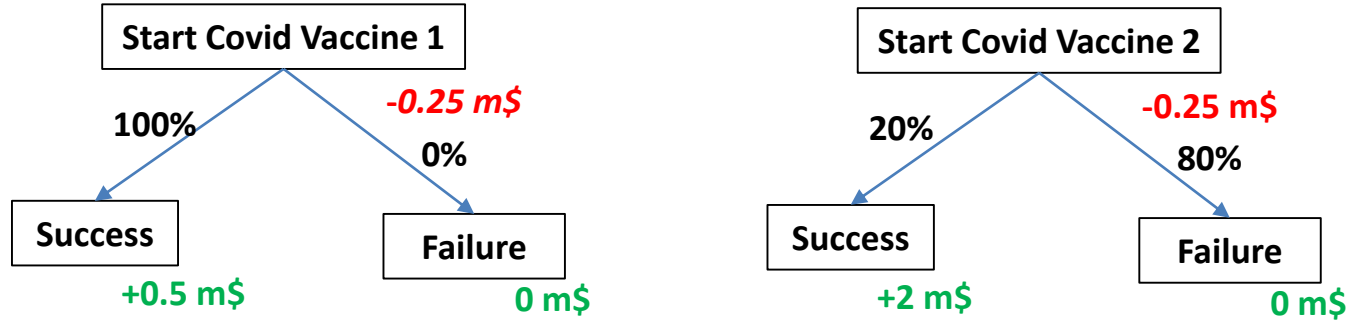
# Models of Experimentation



- **Problem 1:** What if you have to commit to undertaking one project only?
  - Solution: calculate expected value of each project (Vac1:  $0.5 \cdot 1 - 0.25 = \mathbf{0.25}$ , Vac2:  $2 \cdot 0.2 - 0.25 = \mathbf{0.15}$ ) and choose **Vac1** (simple!)
- **Problem 2:** But what if you can undertake both projects sequentially?
  - Strategy 1: Undertake Vac1, then undertake Vac2: Exp. payoff= $\mathbf{0.3}$
  - Strategy 2: Try Vac2 first, then undertake Vac1 if Vac2 fails: Exp. payoff= $2 \cdot 0.2 - 0.25 + 0.8 \cdot (-0.25 + 0.5 \cdot 1) = \mathbf{0.35}$



# Models of Experimentation



- **Problem 1:** What if you have to commit to undertaking one project only?
  - Solution: calculate expected value of each project (Vac1:  $0.5 \cdot 1 - 0.25 = \mathbf{0.25}$ , Vac2:  $2 \cdot 0.2 - 0.25 = \mathbf{0.15}$ ) and choose **Vac1** (simple!)
- **Problem 2:** But what if you can undertake both projects sequentially?
  - Strategy 1: Undertake Vac1, then undertake Vac2: Exp. payoff= $\mathbf{0.3}$
  - Strategy 2: Try Vac2 first, then undertake Vac1 if Vac2 fails: Exp. payoff= $2 \cdot 0.2 - 0.25 + 0.8 \cdot (-0.25 + 0.5 \cdot 1) = \mathbf{0.35}$  -> **hence it makes sense to try Vac2 first!**





# Models of Experimentation

---

- The exploration vs exploitation dilemma:
  - “The [classical] problem where one faces the choice between taking actions which yield immediate reward and taking actions (such as acquiring information, or preparing the grounds) whose benefits will only come by later” (Whittle 1980)
- In our problem, if we tried Covid Vaccine 1 first, **we learn nothing**
- The relatively unpromising Covid Vaccine 2 has a high **option value**: learning whether it works is worth a lot.



# Models of Experimentation

---

- In **Problem 1**, we summarized all relevant information with one value per project that was calculated independently of all other projects (“Expected Value” **EV**)



# Models of Experimentation

---

- In **Problem 1**, we summarized all relevant information with one value per project that was calculated independently of all other projects (“Expected Value” **EV**)
- For **Problem 2**, the same is possible, using the “Reservation Value” **RV**, defined as

$$\int_{RV_i}^{+inf} (x - RV_i) dF_i(x) = c_i$$



# Models of Experimentation

---

- In **Problem 1**, we summarized all relevant information with one value per project that was calculated independently of all other projects (“Expected Value” **EV**)
- For **Problem 2**, the same is possible, using the “Reservation Value” **RV**, defined as

$$\int_{RVi}^{+inf} (x - RVi) dFi(x) = ci$$

**Definition in words:** The certain reward **RVi** so that you are **indifferent** between

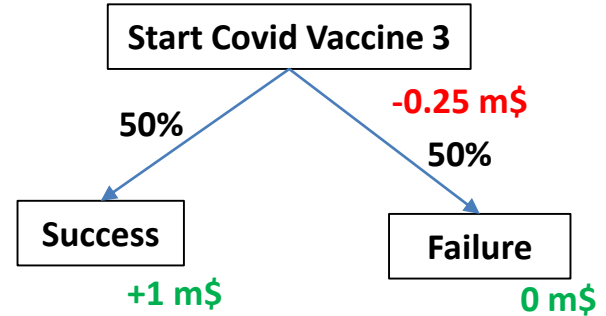
- **accepting RVi**(and not opening box i)
- and opening box i and **accepting max(RVi,BVi)**, where BVi is the reward contained in box i (note that you bear the cost of opening the box).



# Models of Experimentation

## Example: At home

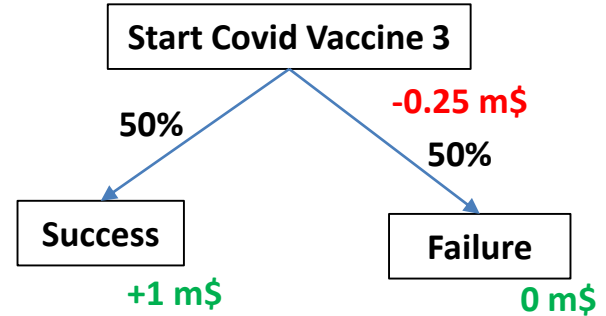
- Suppose that have a certain (proven) vaccine of value **0.8 m\$** (that you developed earlier).
- Is research on Covid Vaccine 3 still worthwhile?



# Models of Experimentation

## Example: At home

- Suppose that we have a certain (proven) vaccine of value **0.8 m\$** (that you developed earlier).
- Is research on Covid Vaccine 3 still worthwhile? If the project fails, we fall back on the old vaccine. Hence, we undertake the project only if:



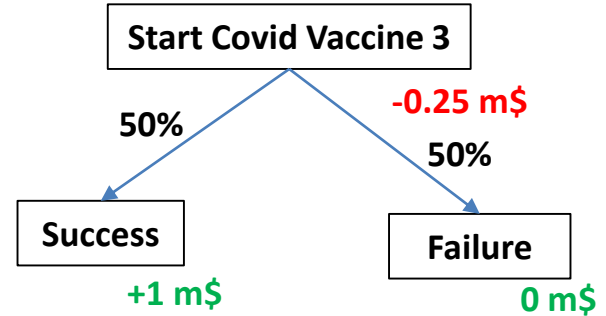
$$0.5 * 1 + 0.5 * 0.8 - 0.25 \geq 0.8$$
$$0.5 * (1 - 0.8) \geq 0.25$$

We only do it if the improvement justifies the cost!

# Models of Experimentation

## Example: At home

- Suppose that we have a certain (proven) vaccine of value **0.8 m\$** (that you developed earlier).
- Is research on Covid Vaccine 3 still worthwhile? If the project fails, we fall back on the old vaccine. Hence, we undertake the project only if:



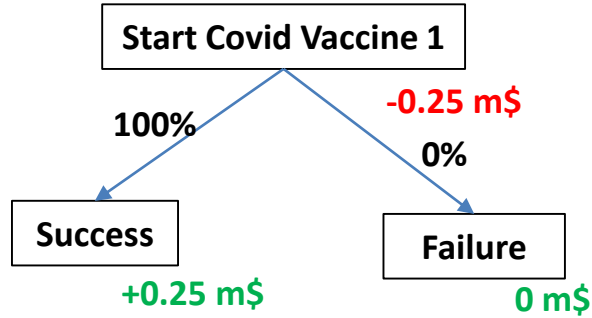
$$0.5 * 1 + 0.5 * 0.8 - 0.25 \geq 0.8$$
$$0.5 * (1 - 0.8) \geq 0.25$$

We only do it if the improvement justifies the cost!

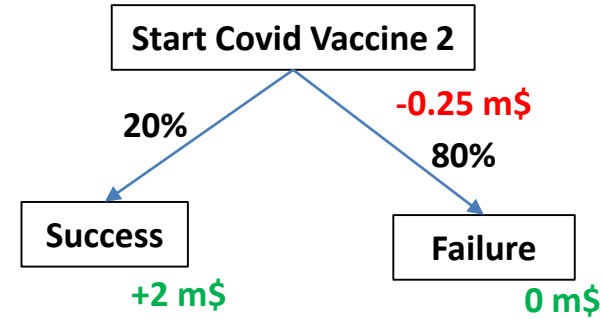
However, there is a cutoff-value for the proven vaccine of **0.5 m\$** that makes this inequality hold with equality, which is the **Reservation Value**



# Models of Experimentation



$$(0.5 - RV1) * 1 = 0.25 \Rightarrow RV1 = 0.25$$

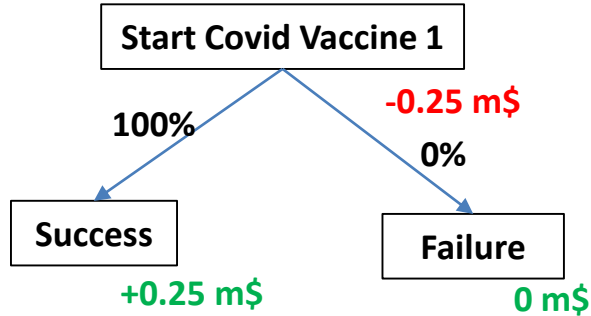


$$(2 - RV2) * 0.2 = 0.25 \Rightarrow RV2 = 0.75$$

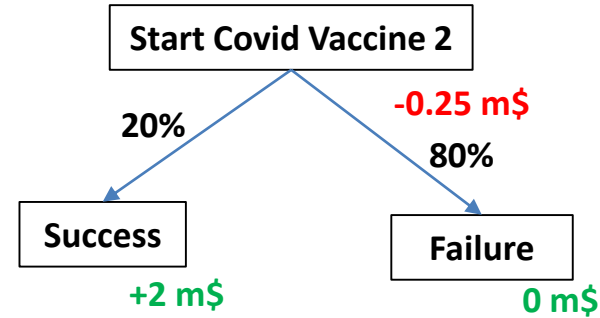




# Models of Experimentation



$$(0.5 - RV1) * 1 = 0.25 \Rightarrow RV1 = 0.25$$



$$(2 - RV2) * 0.2 = 0.25 \Rightarrow RV2 = 0.75$$

## Optimal Strategy:

- Step 1: Order all unopened “boxes” by RV. Select the box with the highest RV and open it.
- Step 2: If the reward from any opened box is greater than the RV of all unopened boxes, stop and take the reward, otherwise continue with Step 1.



# Models of Experimentation

---

## Take-away

- The value that a decision maker derives from innovation is composed of its **expected value** and its **“option value”**
  - Low risk projects may exhibit **high expected values**, but **low option values** (**“Exploitation”**)
  - High risk/high reward projects may often exhibit **low expected values**, but **high option values** (**“Exploration”**)
- Trade-off is managed with index that compares **certain rewards** and **exploring an unknown with the option to fall back** to the certain reward



# Motivating exploration in organizations

---

## The exploration vs exploitation problem

- March (1991):

“A central concern [..] is the relation between the exploration of new possibilities and the exploitation of old certainties. Exploration includes [..] search, risk taking, experimentation, discovery, [..] innovation. Exploitation includes [..] refinement, implementation and execution.”

“[Organizations] that engage in exploration to the exclusion of exploitation suffer the cost of experimentation without gaining its benefits. [Organizations] that engage in exploitation to the exclusion of exploration are trapped in suboptimal equilibria.”



# A quick detour: Multi-armed bandit models

---

## Consider more complex settings

- Suppose you decide on the order of clinical trials investigating the effects of different experimental treatments
- Or you manage a portfolio of research projects that can be stopped/resumed at different times
- Or you optimize portfolio investments in risky startups over time



# A quick detour: Multi-armed bandit models

---

## Consider more complex settings

- Suppose you decide on the order of clinical trials investigating the effects of different experimental treatments
- Or you manage a portfolio of research projects that can be stopped/resumed at different times
- Or you optimize portfolio investments in risky startups over time

## Common features:

- dynamic allocation of resources to different projects
- requires balancing reward maximization based on the knowledge already acquired with attempting new actions to further increase knowledge



# A quick detour: Multi-armed bandit models

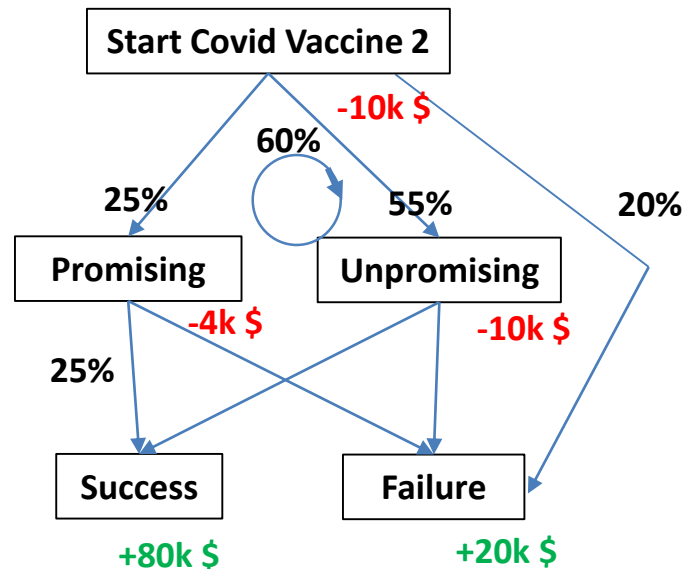
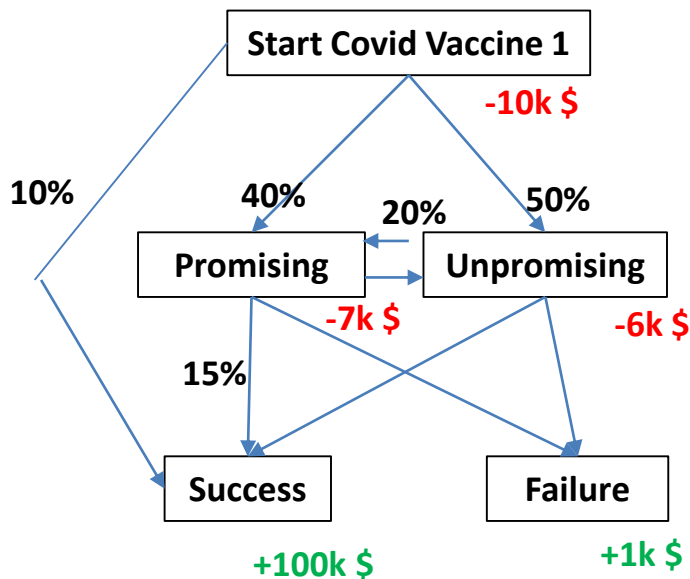
---

- Based on the model used by allied scientist in WW2 for R&D allocation
- First formulated in 1952 by Henry Robbins “The Multi-Armed Bandit Problem”
- Each period, you
  - Choose a bandit
  - Put in the coin
  - Pull the lever
  - The bandit then stochastically changes its “state”, and you potentially collect a reward



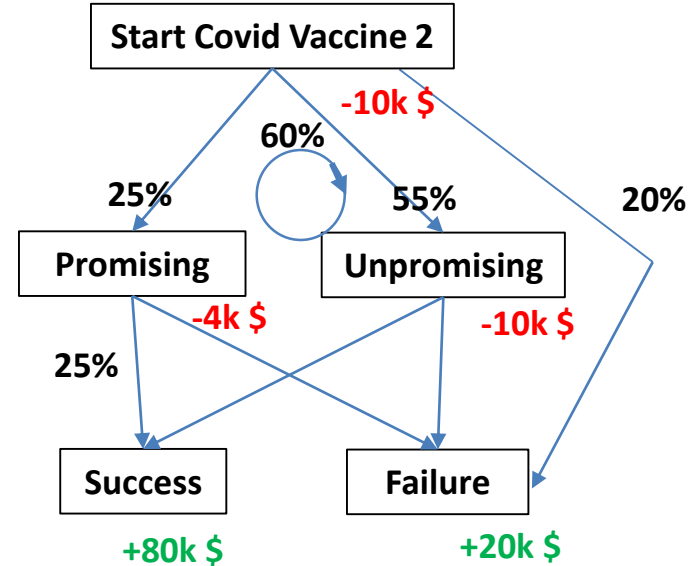
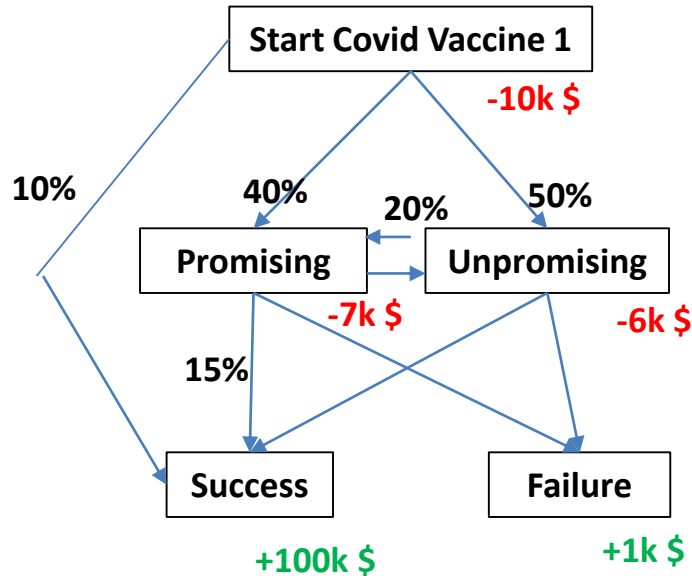
# A quick detour: Multi-armed bandit models

An example: maximize rewards net of incurred costs



# A quick detour: Multi-armed bandit models

An example: maximize rewards net of incurred costs



Possible Strategy: Start Vac 1 -> if it becomes unpromising, start Vac 2 -> if Vac 2 also unpromising, switch back to Vac 1 -> if Vac 1 is a failure, resume Vac 2...





## A quick detour: Multi-armed bandit models

---

- Solution discovered in 1976 by John C. Gittins, the “**Gittins Index**”
  - There exists a value (the “Index”) for each state and each bandit, so that the optimal strategy is to pick the bandit with highest Index at any point in time
  - The calculation of the Gittins Indices **only** requires solving an dynamic optimization problem for each bandit (and state) **independently of all other bandits**.



## A quick detour: Multi-armed bandit models

---

- Solution discovered in 1976 by John C. Gittins, the “**Gittins Index**”
  - There exists a value (the “Index”) for each state and each bandit, so that the optimal strategy is to pick the bandit with highest Index at any point in time
  - The calculation of the Gittins Indices **only** requires solving an dynamic optimization problem for each bandit (and state) **independently of all other bandits**.

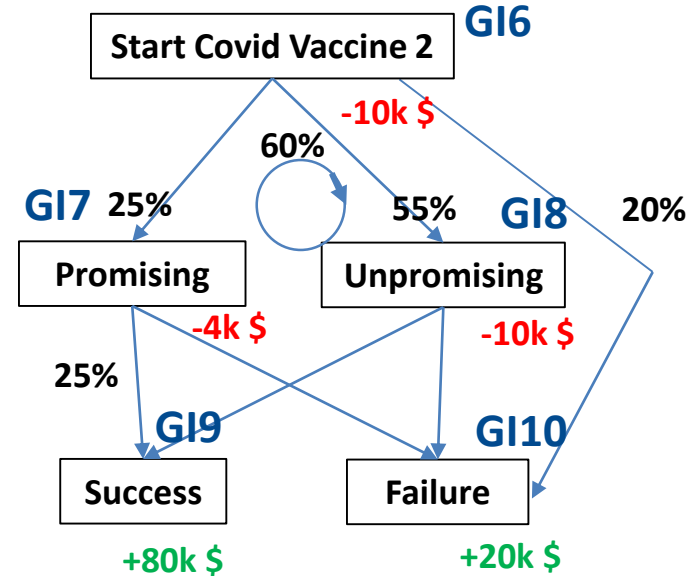
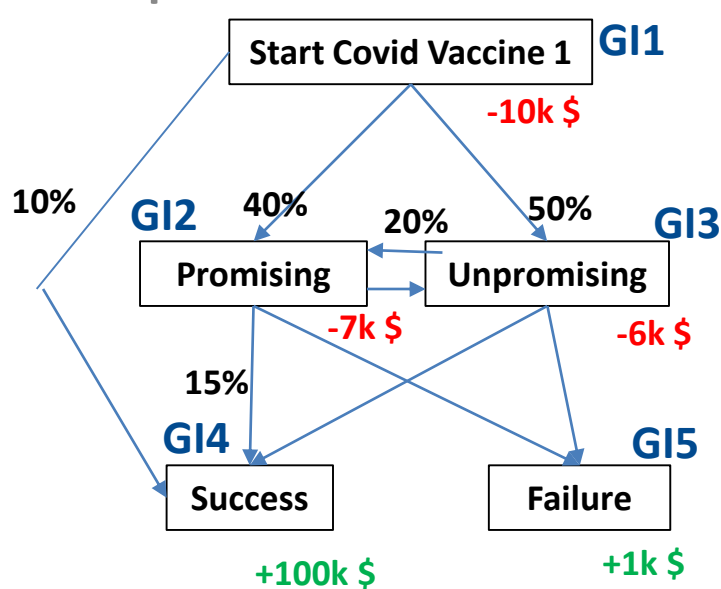
**Definition in words** (Whittle 1980): the certain reward  $GI_i$  so that you are indifferent between

- Accepting  $GI_i$  and “retiring” immediately (stopping the problem) and
- continuing to play for at least one more period and “retire” with (the same) reward  $GI_i$  in the future



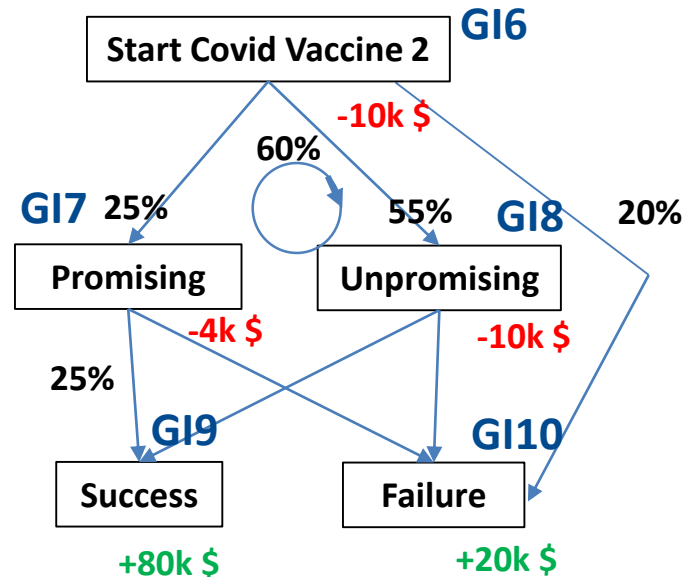
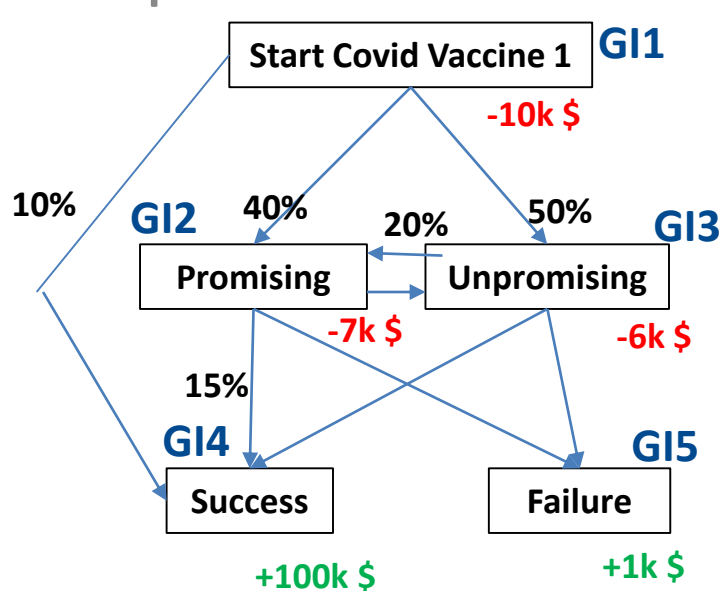
# A quick detour: Multi-armed bandit models

An example: maximize rewards net of incurred costs



# A quick detour: Multi-armed bandit models

An example: maximize rewards net of incurred costs



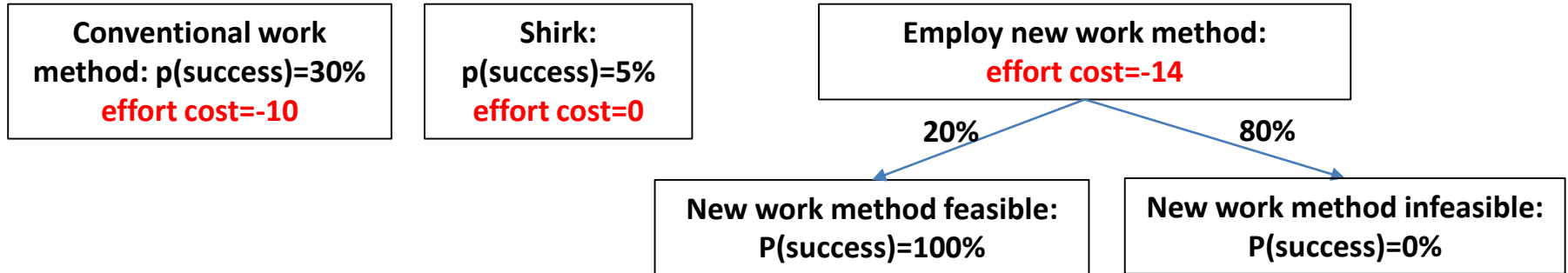
Optimal Strategy: If **GI1** > **GI6**, start Vac1. If Vac1 is unpromising, if **GI6** > **GI3**, start Vac2. If Vac2 is also unpromising, if **GI8** > **GI3** continue with Vac2, otherwise resume with Vac1,...



# Motivating exploration in organizations

## Motivating innovation

- Manso (2011):
  - Suppose you own a company that employs a scientist
  - There are **two periods  $T=1,2$** :
    - Each period, the scientist carries out a task, **which is unobserved**
    - In each period, if the task yields a “success”, the firm earns  **$S=1000\$$** , otherwise 0



# Motivating exploration in organizations

---

## Motivating innovation

**Conventional work  
method:  $p(\text{success})=30\%$   
effort cost=-10**

**Shirk:  
 $p(\text{success})=5\%$   
effort cost=0**

Suppose you want the scientist not to shirk:



# Motivating exploration in organizations

---

## Motivating innovation

**Conventional work**  
method:  $p(\text{success})=30\%$   
**effort cost=-10**

**Shirk:**  
 $p(\text{success})=5\%$   
**effort cost=0**

Suppose you want the scientist not to shirk:

- Standard solution: pay a success-dependent premium (e.g. a share of the profit  $\pi_{1,2} * S$ ) so that in both periods  $T=1$  and  $T=2$ ,



# Motivating exploration in organizations

## Motivating innovation

Conventional work  
method:  $p(\text{success})=30\%$   
**effort cost=-10**

Shirk:  
 $p(\text{success})=5\%$   
**effort cost=0**

Suppose you want the scientist not to shirk:

- Standard solution: pay a success-dependent premium (e.g. a share of the profit  $\pi_{1,2} * S$ ) so that in both periods  $T=1$  and  $T=2$ ,

Scientist's payoff conv. work both periods  $\geq$  Scientist's payoff shirk both periods

$$(0.3 * S * \pi_{1,2} - 10) * 2 \geq (0.05 * S * \pi_{1,2}) * 2$$
$$\pi_{1,2} \geq 0.04$$





# Motivating exploration in organizations

---

## Motivating innovation

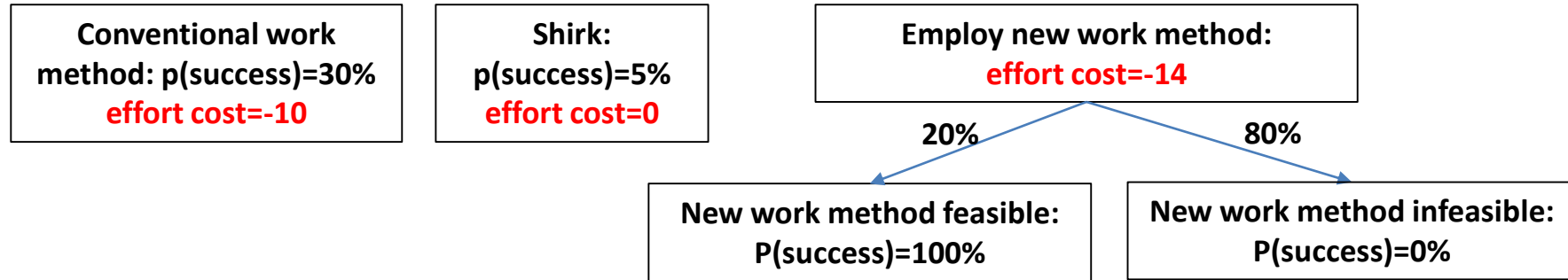
Discussion:

- A standard **“Pay for Performance” scheme** that pays the scientist a 4% premium (share of profits) period-per-period solves the incentive problem
- It ensures that the scientists prefers the conventional work method in each period over the “shirking” option
- The expected profit of the firm under the “Pay for Performance” scheme is
$$0.3 * S * (1 - \pi^{PP}) + 0.3 * S * (1 - \pi^{PP}) = 576 \$$$



# Motivating exploration in organizations

## Motivating innovation



We consider the following “exploration” strategy:

- In  $T=1$ , employ the new work method
- In  $T=2$ , if the new work method is feasible, employ it again. Otherwise, switch to the conventional work method in  $T=2$ .



# Motivating exploration in organizations

---

- Scientist's payoff under exploration:

$$0.2 S\pi_1 - 14 + 0.2(S\pi_2 - 14) + 0.8(0.3 S\pi_2 - 10)$$

- Expected Firm profits under exploration:

$$0.2 S(1 - \pi_1) + 0.2 S(1 - \pi_2) + 0.8 0.3 S(1 - \pi_2)$$

## Verify at home !:

- Under any “pay-for-performance” scheme with a premium in the range of 4% to 12% , the scientist prefers to employ the conventional work method (in both periods) over the exploration strategy
- With a “pay-for-performance” scheme of a premium of more than 12%, exploration is preferable to the scientist, but the profit of the firm drops below the profit level that it could achieve with a “pay-for-performance” scheme of 4% (and having the scientist not explore the new work method), which was 576\$



# Motivating exploration in organizations

---

- Why is period-per-period “pay-for-performance” bad for motivating exploration?
    - In the first period, the expected payoff of the new work method is lower than the conventional method (a hallmark property of exploration problems)
    - Benefits of exploration accrue later, but “pay-for-performance” rewards short-term and long-term success equally
- ➡ **“Short-termism”**: Short-term rewards hinder the exploration of a-priori unpromising approaches



# Motivating exploration in organizations

---

- What is the **solution**?
  - Consider the “**deferred compensation**” scheme  $\pi_1 = 0 \%$  and  $\pi_2 = 12 \%$
  - It ignores success and failures in period  $T=1$  and only considers the later period  $T=2$
  - By focusing on “long-term” outcomes only, it preserves the incentives to explore.

## Verify at home!:

- Under the “deferred compensation” scheme with a premium of 12%, the exploration strategy is preferred by the scientist over:
  - Employing the conventional work method in both periods
  - Shirking in  $T=1$  and employing the conventional work method in  $T=2$
- The profit of the firm is higher than under the 4% pay-for-performance scheme (of 576\$)



# Short-termism

---

## X. Tian, T. Y. Wang (2014, RFS): "Tolerance for Failure and Corporate Innovation"

- They measure the average investment duration of Venture Capital firms into startups that eventually failed (on average: 3.25 years) as a measure of **tolerance for failure**
- Especially experienced (i.e. older) VC firms exhibit failure tolerance



# Short-termism

## X. Tian, T. Y. Wang (2014, RFS): "Tolerance for Failure and Corporate Innovation"

- They measure the average investment duration of Venture Capital firms into startups that eventually failed (on average: 3.25 years) as a measure of **tolerance for failure**
- Especially experienced (i.e. older) VC firms exhibit failure tolerance
- They find that if a **startup is backed by a failure-tolerant VC firm** (at the 75<sup>th</sup> percentile in tolerance), it is **predicted to have 39% more patents** than a startup backed by a failure-intolerant VC firm(at the 25<sup>th</sup> percentile)



# Short-termism

P. Azoulay, J.S. Graff Zivin, G. Manso (2011, RAND): "Incentives and creativity: evidence from academic life sciences"

- They compare the publication output of researchers funded through
  - long-term grants (typ. 10 years) with **full freedom**
  - short-term grants renewed every 3-5 years with **review of accomplishments**





# Short-termism

P. Azoulay, J.S. Graff Zivin, G. Manso (2011, RAND): "Incentives and creativity: evidence from academic life sciences"

- They compare the publication output of researchers funded through
  - long-term grants (typ. 10 years) with **full freedom**
  - short-term grants renewed every 3-5 years with **review of accomplishments**
- Researchers on long-term grants do **not produce more average publications**, but they produce
  - **twice as many "hits"** that are more heavily cited and by a broader set of related subfields
  - And more publications that introduce **"new technical keywords"**



# References

---

- Weitzman, M.L. (1979): Optimal search for the best alternative. *Econometrica*, 47, p. 641-654
- Gittins, J.C. (1979): Bandit Processes and Dynamic Allocation Indices. *Journal of the Royal Statistical Society. Series B (Methodological)*. 41, p. 148-177
- March, J.G. (1991): Exploration and Exploitation in Organizational Learning. *Organization Science*, 2, p. 71-87
- Manso, G. (2011): Motivating Innovation. *The Journal of Finance*, 66, p. 1823-1860



# References

---

- Azoulay, P., Graff Zivin, J.S. & Manso, G. (2011): Incentives and creativity: evidence from the academic life sciences. *RAND Journal of Economics*, 42, p. 527-554
- Tian, X. & Wang, T.Y. (2014): Tolerance for Failure and Corporate Innovation. *The Review of Financial Studies*, 27, p. 211-255

