Terminating Information Search

Stefan Appelhoff

November 16, 2016

Contents

List of Figures

1 Introduction

As agents in different environments, humans often find themselves in situations that require decisions based on limited knowledge about the available choice options. The topic of the present project is the process, how humans selectively sample choice options

to arrive at a more refined knowledge about the properties of the options. Specifically, the interest lies in self-directed information search in sequential decision making problems. A short example illustrates the topic further:

> Imagine that your laptop is broken and you want to acquire a new one. In the process, you may sequentially sample several available models and their respective properties such as prices, specifications, and reviews. To avoid indefinite browsing, you will need to terminate your search at some point and select a new laptop - most likely without having obtained full certainty about the best option.

For the project described here, the aforementioned example will be transformed into an experimental paradigm that allows for the examination of behavioral and neural aspects of choice in a controled manner. Similar to the real life example, the experimental paradigm exhibits the following characteristics:

- There is an environment in which several options exist that will yield outcomes upon selection.
- There exists limited knowledge about the options and the frequencies of their associated outcomes.
- There will be the opportunity to obtain information about the options through a sampling of the options.
- The overall goal is to maximize the positive outcomes obtained from the options in the long run.

In the paradigm, participants will make sequential decisions about which option to sample. In these trial to trial decisions, the question of interest to this project lies in how participants set their termination criteria. I.e., when do participants decide to terminate the information search in one option and switch to search another option? When do participants decide to stop sampling altogether and decide for one final option to be optimal based on current knowledge? The remainder of this document will be a concise formulation of the research questions followed by an outline of the experimental paradigm. Finally, the implementation of the experimental paradigm on a computer program will be described.

2 Research Questions

1. How are sampling efforts sequentially allocated to the different options?
2. How do we arrive at a decision to terminate the information search?
3. What is the quantitative link between neural signals obtained by EEG with the behavioral data of participants performing the experimental paradigm.

3 The Experimental Paradigm

The proposed experiment will be based on a combination of two different paradigms that are well known in the literature.

1. The n-armed Bandit Paradigm
2. The Sampling Paradigm

In the following, these two paradigms will be descibed in detail. Figure 1 provides an overview of the paradigms. Afterwards, the combination of both paradigms for the present experiment will be outlined.

3.1 The n-armed Bandit Paradigm

The n-armed bandit paradigm was first formalized by Robbins in 1952 [5]. It has since been the subject of interest in a variety of fields, see [1] for a review. Generally, participants are presented with n action that represent initially unknown outcome distributions with a set number of outcomes. Upon selection of an action, an outcome is yielded from the distribution and presented to the participant as a feedback. With this gained knowledge, the participant may continue with selecting the same or a different action.

The bandit paradigm offers a concise framework to study the exploration exploitation tradeoff common to the field of reinforcement learning. Here, exploration represents the goal to gather information about actions and their associated outcomes, while exploitation represents the goal to make use of the acquired information. A player who only explores will never exploit upon the accumulated knowledge, while a player who only exploits will most likely select suboptimal actions based on the inadequate knowledge about all actions. Hertwig & Erev label the bandit paradigm as Partial-Feedback-Paradigm [4].

3.2 The Sampling Paradigm

The sampling paradigm is similar to the bandit paradigm. The main difference is that exploration and exploitation are separated. This is achieved through permitting a participant to sample all actions without consequences for as long as the participant wishes. This stage represents purely the exploration, because the outcomes do not count towards a payoff. Once the participant decides to terminate the sampling process, there is one last opportunity to select any action. For this last choice, the outcome will affect the payoff and thus, this stage represents purely the exploitation. For more information, see [2–4]

3.3 Combination of the Two Paradigms

In the present project, we will apply a paradigm that combines the aforementioned bandit and sampling paradigms into a single experimental flow. We will further introduce another factor, namely the mode of interaction of the participant with the paradigm. Each of the factors will have two levels. In brief, we are thus talking about a 2x2 factorial design of an experiment as visualized in figure 2. The two factors are:

- Paradigm condition
- Interaction condition

3.3.1 Paradigm condition

The paradigm condition is governed by the bandit or sampling paradigm as the two levels respectively. Thus in the bandit condition, participants will perform a bandit task, whereas in the sampling condition, participants will perform a sampling paradigm task. This will allow us to uncover potential differences in behavior when humans are faced with a task that either combines or separates exploration and exploitation into stages.

3.3.2 Interaction condition

The interaction condition is goverened by active or passive interaction respectively. During the active condition, participants will actively select the actions and steer the flow of the paradigm. During the passive condition, participants will merely watch a replay of previous games. This allows us to make a distinction between (a) cognitive and motor involvement in active tasks and (b) neither cognitive nor motor involvement in the passive tasks. During the passive tasks there is no cognitive involvement, because participants are not solving or learning the underlying decision paradigm.

3.3.3 Orthogonal task

To make sure that participants pay attention even during the passive conditions, we introduce an orthogonal task to the 2x2 experimental design. Here, participants will have to react to a stimulus of a different color by a quick button press. If the reaction times are too slow, we can assume that the participants did not pay sufficient attention.

4 Computerized Implementation of the Experiment

1. **The GUI**
   At the beginning of the experiment, a graphical user interface will pop up, inquiring about the following information:
   - Subject ID

- The subject ID will have at least three digits so that participant one will be '001', participant 2 will be '002' and so on.

- Stimulus color
  - The stimulus color can be set to either 'blue' or 'red', this entry determines, which color the winning stimulus will be. The losing stimulus will be set to the opposite color (e.g., 'blue', if the selection was 'red').

- Starting condition
  - Finally, the starting condition can be set to either 'sp', which stands for sampling paradigm, or 'pfp', which stands for partial feedback paradigm (i.e., bandit). The selected option will be the initial condition of the experiment starting with 'active' as the second factor. This will be followed by a replay of the selected option (i.e., passive condition), before the move goes to the remaining condition, again in the order active - passive.

2. **Instructions**
   Right after the initial GUI, there will be a set of initial instructions. I.e., the outline of the overall paradigm. Furthermore there will be more instructions before each upcoming condition.

3. **Active PFP Conditions**
   If we have 100 trials for this condition, there will be about 4 'games' with 25 trials each. For each game, the lottery locations will be shuffled, so each game presents a new learning situation. See figure 3 for an overview.
   - 'The lotteries have been shuffled' screen [1 sec.]
   - Trial counter current_trial / overall_trials [1 sec.]
   - Fixation cross
   - Reaction by participant [left] or [right] key to mimic selection of left or right lottery
   - Fixation cross screen: delay before outcome of choice is presented [1 sec.]
   - Feedback displayed in form of stimulus (see figure 5) [1 sec.]
   - Continue with next trial counter until current_trial == overal_trials
   - Show the payoff for the currently finished game [1 sec.]
   - Ask the participant which lottery was deemed more profitable.
   - Then, next game of PFP. Start with 'lotteries shuffled' screen

4. **Passive PFP Condition**

   - Replay of the Active PFP Condition

- Take all timings as they were recorded before, and only interfere, when the timings are inadequately high or low. In these cases, we select a timing at random from the interval [0, 1] as implemented in matlab by: rand+rand.
- Skip the question which lottery was deemed more profitable

5. **Active SP Condition**
   As opposed to the active PFP condition, we do not separate our overall 100 trials into distinct games. Instead, we will shuffle the lotteries after each 'choice' stage. Thus, there can be potentially many or few novel learning situations, depending on how often a participants picks 'choice' over 'continue to sample'. See figure 4 for an overview.

   - 'The lotteries have been shuffled' screen [1 sec.]
   - Trial counter current_trial [1 sec.]
   - Fixation cross
   - Reaction by participant [left] or [right] key to mimic selection of left or right lottery
   - Fixation cross screen: delay before outcome of choice is presented [1 sec.]
   - Feedback displayed in form of stimulus (see figure 5) [1 sec.]
   - Question whether participant would like to:
     - Make a choice
     - Continue to sample
   - display choosen answer [0.7 sec.]
   - If 'continue to sample' was selected, continue with next trial counter else, present a choice screen.
   - Show the payoff for the currently finished game [1 sec.]
   - Then, continue with SP. Start with 'lotteries shuffled' screen

6. **Passive SP Condition**

7. - Replay of active SP condition
   - Timings are taken from active SP condition, unless they are inadequate (see also passive PFP condition)

8. **Orthogonal Task**
   Throughout all four possible conditions, there will be a set chance (5%) that instead of a blue or red stimulus, a green stimulus (5) will occur. As soon as that stimulus is shown, participants ought to press the [space] key.

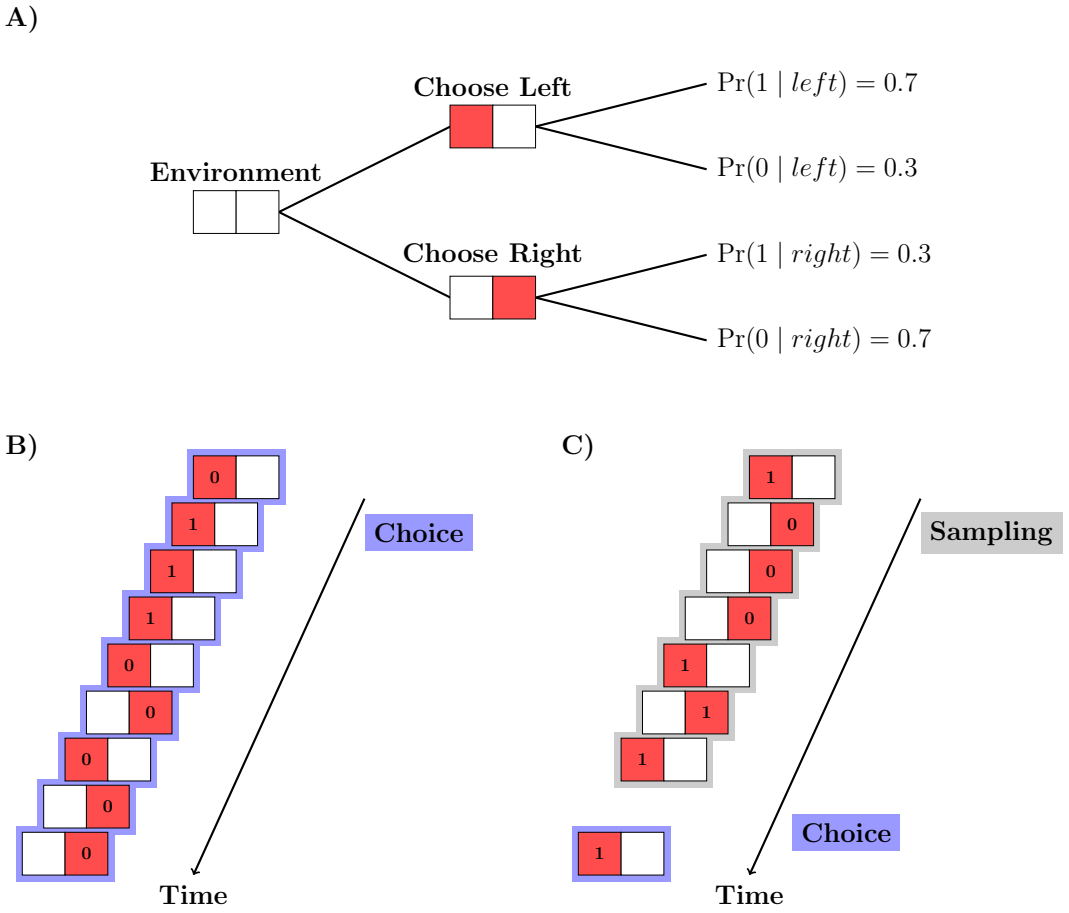

Figure 1: A) Depiction of a two-armed bandit. Note that the number of options in the environment can be extended arbitrarily to form an n-armed bandit. Each option contains its own probability mass function (pmf) of outcomes. These pmfs are unknown to a subject performing the bandit task. **B)** A typical run of trials in a bandit task. A subject sequentially chooses among options and is provided with feedback. Each trial represents a exploration-exploitation problem. **C)** A typical run of trials in the sampling paradigm. A subject is allowed to sample the available options without an impact of the outcomes on the final payoff. At some point, the subject can decide to terminate sampling and choose one of the options, which will impact the final payoff. In the sampling paradigm, exploration and exploitation are thus distinct.

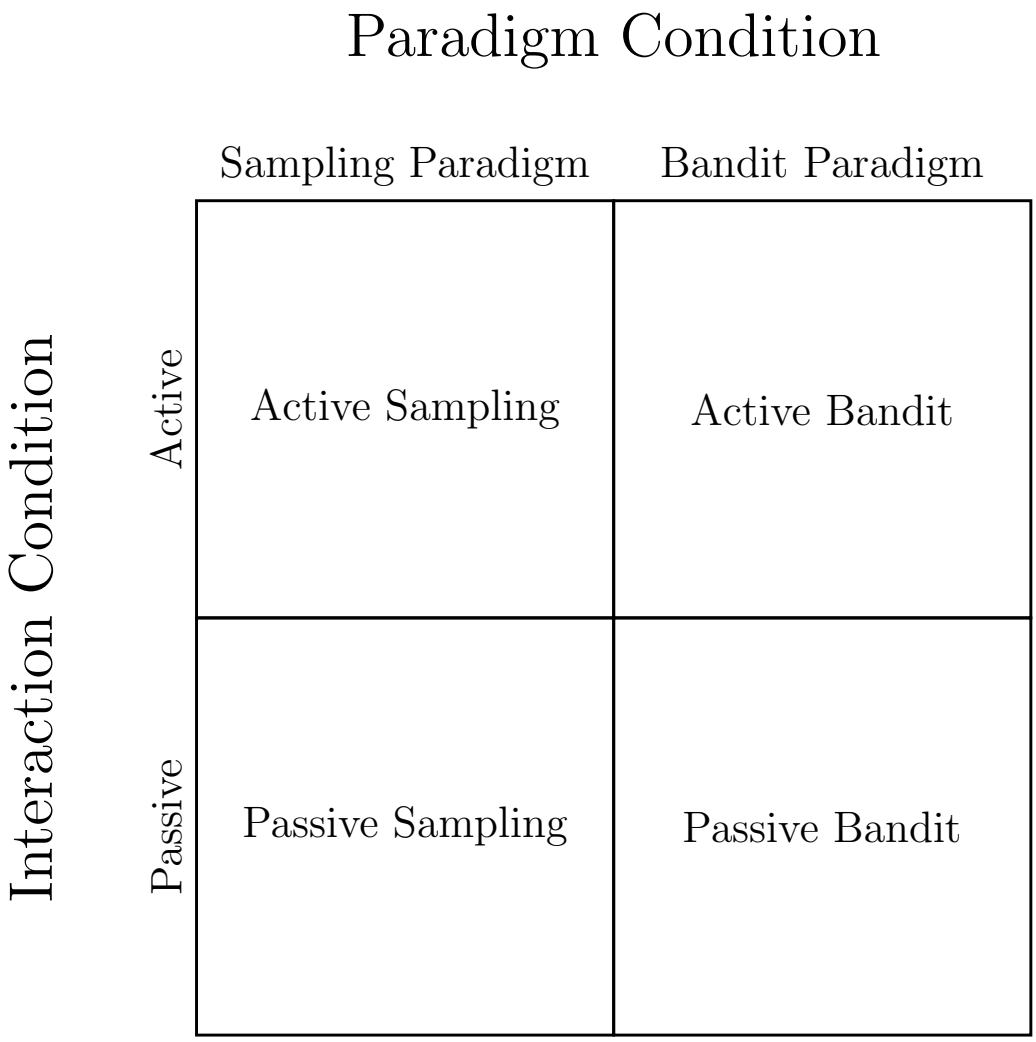

Figure 2: The 2x2 factorial design of the experiment.

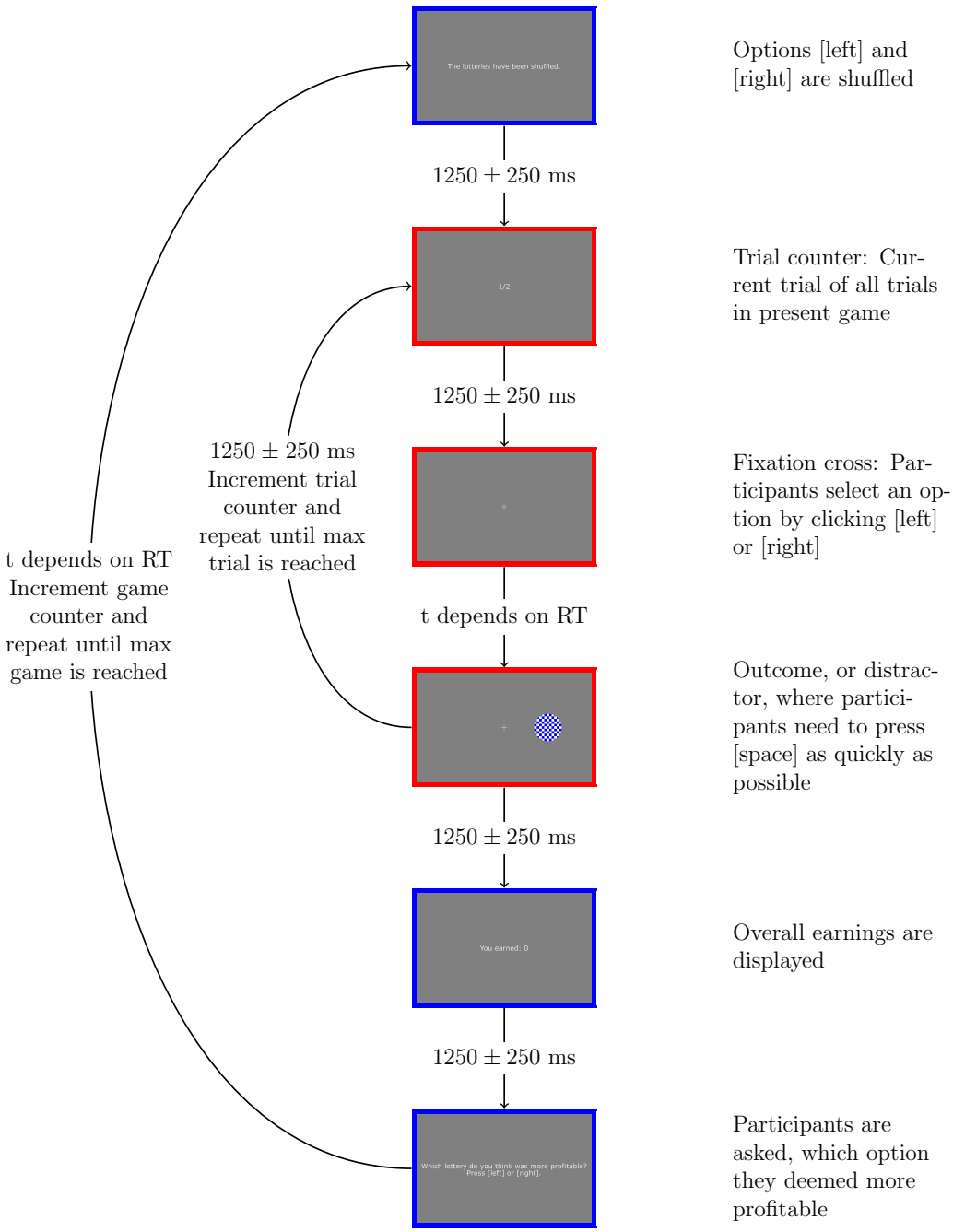

Figure 3: Experimental flow of the bandit paradigm. Red colors indicate the loop where participants explore and exploit there options. Once all trials have been spent within the red loop, a transition to the blue loop occurs, resetting the environment for a new game and leading to the red loop again, until all games are exhausted. Note: RT=Reaction time, ms=miliseconds, t=time

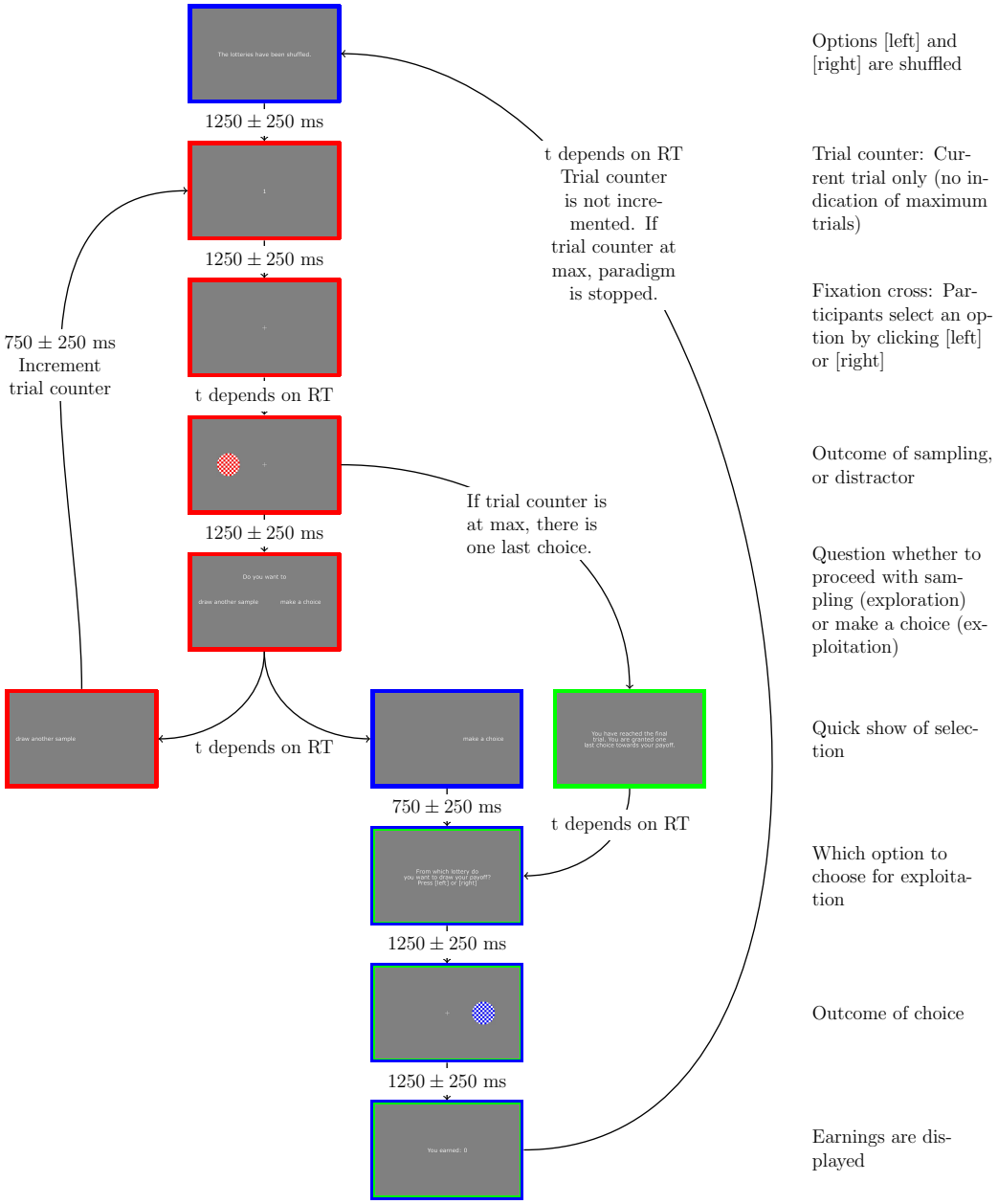

Figure 4: Experimental flow of the sampling paradigm. Red colors indicate the sampling loop, from which one can transition to the blue choice loop, leading back to the sampling loop. Once the maximum of trials has been reached, a transition to the green loop for a last choice occurs. Note: RT=Reaction time, ms=miliseconds, t=time.

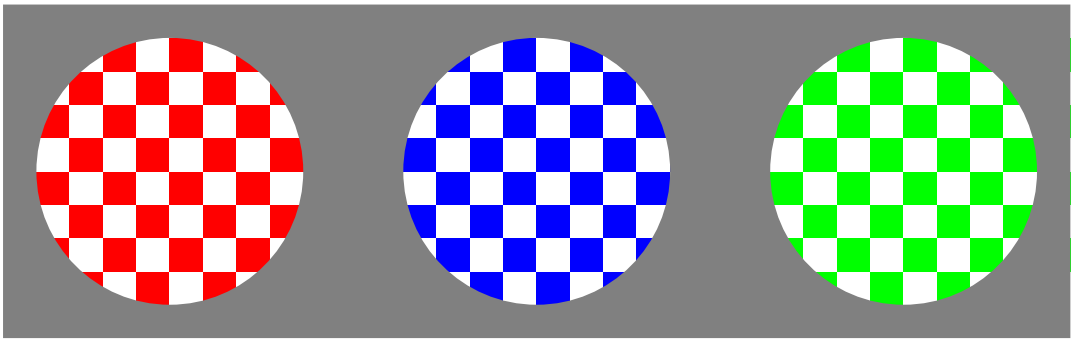

Figure 5: Three stimuli against a background as used in the experiment . The red and blue stimuli represent either a win or lose outcome. The green stimulus represents a distractor stimulus.

5 References

[1] Donald A Berry and Bert Fristedt. *Bandit problems: sequential allocation of experiments (Monographs on statistics and applied probability)*. Springer, 1985.

[2] Ido Erev and Alvin E. Roth. "Maximization, Learning and Economic Behavior". In: *Proceedings of the National Academy Sciences of the USA* 111.3 (2014), pp. 10818–10825. DOI: 10.1037/e573552014-045. URL: http://dx.doi.org/10.1037/e573552014-045.

[3] Ralph Hertwig. "The psychology and rationality of decisions from experience". In: *Synthese* 187.1 (Oct. 2011), pp. 269–292. DOI: 10.1007/s11229-011-0024-4. URL: http://dx.doi.org/10.1007/s11229-011-0024-4.

[4] Ralph Hertwig and Ido Erev. "The description–experience gap in risky choice". In: *Trends in Cognitive Sciences* 13.12 (Dec. 2009), pp. 517–523. DOI: 10.1016/j.tics.2009.09.004. URL: http://dx.doi.org/10.1016/j.tics.2009.09.004.

[5] Herbert Robbins. "Some Aspects Of The Sequential Design Of Experiments". In: *Bulletin of the American Mathematical Society* 58 (Sept. 1952), pp. 527–535.