Probability in the Infinite Hat Game

Submission for the *2018 MITx Philosophy Award*

Pascal Bachor*

Part I: The Prisoner's Perspective

Let $\{P_i\}_{i\in\mathbb{N}}$ be a set of *prisoners*. Each prisoner P_i will randomly, by toss of a fair coin, be assigned a *hat color* $\alpha_i \in \{0,1\}$, where $\alpha \in \mathcal{A}$ and $\mathcal{A} \coloneqq \mathbb{Z}_2^{\mathbb{N}}$ is the set of possible hat assignments[1]. We will sometimes refer to α as the *actual assignment*. In the *infinite hat game* or *Bacon's Puzzle* the prisoners will be placed on a line such that each prisoner P_i can see all prisoners $\{P_j \mid j > i\}$ (and their hat color). They shall now simultaneously declare (a guess of) their own hat color. We say a prisoner *wins* if they correctly declare the color of their hat, else they *lose*.

For $a, b \in \mathcal{A}$ we say $a \sim b$ iff a and b differ at at most finitely many places. It is easy to see that $\sim$ is an equivalence relation. Let $\mathfrak{c} : \mathcal{A} \to \mathcal{A}$ be a function that assigns to each element a unique representative from their equivalence class, that means $a \sim b \implies \mathfrak{c}(a) = \mathfrak{c}(b)$. Let $E_r \coloneqq \{a \in \mathcal{A} \mid a \sim r\}$. Assuming the prisoners know $\mathfrak{c}$, we can construct a strategy for them that assures that at most finitely many of them will declare the wrong color: Each prisoner P_i identifies the equivalence class $E_{\mathfrak{c}(\alpha)}$ of α and declares $\mathfrak{c}(\alpha)_i$ (this is possible because P_i sees all but finitely many places of α). We will refer to this strategy as *the **S**trategy*.

In this paper we will treat the question *What is the probability for a particular prisoner to declare correctly?* and what role it plays in our treatment of the puzzle. For simplification we will refer to this probability also as *the **P**robability*.

Let us first look into how Prof. Rayo treats the Probability. At the end of the section in which the puzzle is introduced he states the following. "I suspect [..] one's probability of success is best thought of ill-defined.". The puzzle is later revisited at the end of the chapter *Non-Measurable sets*[2]. There he assumes the perspective of some prisoner *during the game*, that is after the hat colors have been assigned but before declarations are made.

*bachorp@informatik.uni-freiburg.de

[1] We use $\mathbb{Z}_2 = \{0,1\}$ to make use of the modular addition $a +_{\mathbb{Z}_2} b \coloneqq a + b \mod 2$.

[2] I would not be able to state what's following without a precise explanation Prof. Rayo kindly gave to me personally.

For that, we will now fix some prisoner P_k and define $\mathfrak{W}_i \coloneqq \{a \in \mathcal{A} \mid a_i = \mathfrak{c}(a)_i\}$, the set of assignments for which P_i will win.

Next, he argues that the probability for P_k to win is equal to the proportion of winning assignments inside the equivalence class of the actual assignment $E_{\mathfrak{c}(\alpha)}$ (which P_k is able to identify); Or equivalently, the probability of the actual assignment to be in $\mathfrak{W}_k$ given that it is in $E_{\mathfrak{c}(\alpha)}$.

Until this point I very much agree with his line of reasoning. It is, in my opinion, a good idea to restrict ourselves to a subset of possible assignments (in this case $E_{\mathfrak{c}(\alpha)}$) when trying to make sense of the Probability. And in fact we will continue in this manner below.

Now though, he proceeds by making what he previously called a suspicion a claim. He states "I claim that there is no good answer to what that probability is.". He does not give any evidence supporting this claim (nor does he pretend to have any) so we are left with his belief.

Claim (Rayo)**.** *The probability for some prisoner to declare correctly should be considered ill-defined.*

When assuming P_k's perspective, I think Prof. Rayo's argumentation is not quite followed through. Namely, the assignments that agree with the information we are given are not $E_{\mathfrak{c}(\alpha)}$ but only $T_{k,\alpha} \coloneqq \{a \in \mathcal{A} \mid a_{k+1+i} = \alpha_{k+1+i} \text{ for all } i \in \mathbb{N}\}$, the assignments that have the infinite tail seen by P_k. Thus, we would investigate the probability of $\mathfrak{W}_k$ given $T_{k,\alpha}$.

All the $k+1$ unknown places of the assignments in $T_{k,\alpha}$ are determined by independent coin tosses. Thus, the probability for each of these assignments to be the actual assignment should be considered equal to $\frac{1}{2^{k+1}}$. As the declaration we will make is the same for all $T_{k,\alpha}$ and exactly half of the 2^{k+1} assignments have a 0 (resp. a 1) at position k, exactly 2^k assignments will agree with the declaration (s.t. we win). Using *Additivity*, the probability for the assignment to be of this kind would then be considered equal to $\frac{2^k}{2^{k+1}} = \frac{1}{2}$.

This, to me, is a convincing argument for why the probability to win from P_k's perspective should be considered 50%. Therefore, I will refuse Prof. Rayo's claim and claim the following instead.

Claim. *The probability for some prisoner to declare correctly should be considered equal to* $\frac{1}{2}$.

Part II: The Mathematician's Perspective

We will now take a broader perspective and consider the situation *before the game starts*, that is before any hat colors have been assigned. We first determine some properties that we would like a reasonable probability measure over $\mathcal{A}$ to have and then investigate how this affects the probability of the event $\mathfrak{W}_k$.

Definition. $\mu : \mathcal{P}(\mathcal{A}) \to [0,1]$ *is a* measure *if for* $A, B \subseteq \mathcal{A}$ *with* $A \cap B = \emptyset$ *it holds*

(i) $\mu(\mathcal{A}) = 1$ *and*

(ii) $\mu(A \cup B) = \mu(A) + \mu(B)$.

The reasoning behind these constraints has been discussed in the course.

Definition. *For* $a, b \in \mathcal{A}$ *we have* $(a + b) \in \mathcal{A}$ *with* $(a + b)_i \coloneqq a_i +_{\mathbb{Z}_2} b_i$ *for all* $i \in \mathbb{N}$. $A^{\to d} \coloneqq \{a + d \mid a \in A\}$ *is the* translation of A by d.

Note how this compares to the definition we have seen in the course.

Definition. $\mu : \mathcal{P}(\mathcal{A}) \to [0,1]$ *is* uniform *if for all* $A, B \subseteq \mathcal{A}$ *with* $B = A^{\to d}$ *for some* $d \in \mathcal{A}$ *it holds* $\mu(A) = \mu(B)$.

To see why this would be a reasonable property, let's say the assignment α is constructed using the following procedure: For each P_i we flip a fair coin C_i. If it shows Heads we set $\alpha_i = 0$, else $\alpha_i = 1$. Because the coins are fair reversing this rule for some of the coins (i.e. $\alpha_i = 0$ iff C_i shows Tails) won't change any of the probabilities. Let $B = A^{\to d}$. Suppose we reverse the rules for all coins $\{C_i \mid d_i = 1\}$. Now, for any outcome of the coin tosses an assignment will be in B iff it would be in A under the usual rules. Thus, the probability for an assignment to be in A should be equal to the probability for it to be in B.

Uniformity captures, in a way, the fact that the coin tosses are fair and independent. For example $\mu(\{a \mid a_i = x\} \cap \{a \mid a_{i+1} = 0\}) = \mu(\{a \mid a_i = x\} \cap \{a \mid a_{i+1} = 1\})$. Whatever C_i shows, C_{i+1} will show Heads with the same probability as Tails.

Theorem. *Let* M *be a uniform measure, then* $\mathrm{M}(\mathfrak{W}_k) = \frac{1}{2}$.

Proof. Let $d \in \mathcal{A}$ such that $d_i = 1 :\Leftrightarrow i = k$ and $\mathfrak{L}_k \coloneqq \mathfrak{W}_k^{\to d}$. Note that $\mathfrak{L}_k$ forms the set of assignments for which P_k loses. It holds

$$\mathrm{M}(\mathfrak{W}_k) = \mathrm{M}(\mathfrak{L}_k) \text{ by uniformity,} \tag{1}$$

$$\mathfrak{W}_k \cap \mathfrak{L}_k = \emptyset \text{ and} \tag{2}$$

$$\mathfrak{W}_k \cup \mathfrak{L}_k = \mathcal{A} \text{ as } P_k \text{ wins iff they don't lose.} \tag{3}$$

Using the definition of measure it follows

$$\mathrm{M}(\mathfrak{W}_k) + \mathrm{M}(\mathfrak{L}_k) \overset{(ii),(2)}{=} \mathrm{M}(\mathfrak{W}_k \cup \mathfrak{L}_k) \overset{(3)}{=} \mathrm{M}(\mathcal{A}) \overset{(i)}{=} 1.$$

With (1) follows $\mathrm{M}(\mathfrak{W}_k) = \frac{1}{2}$. □

The Way Out

Finally, we want to asses our view of the puzzle in the context of what we've seen in this paper. The paradoxical situation of Bacon's puzzle boils down to these supposedly contradictory statements[3].

1. Each prisoner has a 50% chance to win.

2. All but finitely many prisoners will win in any case.

If we consider the Probability ill-defined in case the Strategy is implemented, we can reject the first statement and escape the puzzle this way. We have argued against that.

If the prisoners follow the Strategy, we cannot reject the second statement and we have argued for accepting also the first. The way out is, in my opinion, that the two statements only seem contradictory when in fact they are not. One might be tricked into thinking that if the Probability is 50%, using the Strategy would be comparable to each prisoner tossing a coin to decide their declaration. The important difference is, that the declarations the prisoners make with the Strategy are not independent of each other. Thus, the possibility for the prisoners to coordinate in certain ways is not ruled out. Also, one might tend to overestimate the fact that the number of losing prisoners is finite and forget that it is in fact not bounded across multiple instances of the game. Thus, any prisoner has almost no prospect of being part of the infinite tail of prisoners that are guaranteed to win[4].

I hope this paper has not only (I) shown the necessity of accepting both statements but also (II) helped in making sense of how this is possible. The latter, amongst other things, by showing how a reasonable probability measure could look like in an environment where both statements hold[5].

[3] As laid out in the subsection *Puzzle Solved?*.

[4] Similar observations are made by Prof. Rayo.

[5] A uniform measure assigns a probability of $\frac{1}{2}$ to all events $\mathfrak{W}_i$, while all assignments are part of all but finitely many of the $\mathfrak{W}_i$.