Notes on Bacon's Puzzle

Agustín Rayo, MIT

A. Bacon's puzzle

1. The Puzzle

Andrew Bacon. A paradox for supertask decision makers. *Philosophical Studies*, 153(2):307–311, 2011

- There is an ω-sequence of people: $P_1, P_2, P_3 \ldots$. Each person is wearing a hat. For each k, a coin was flipped to determine whether to give P_k a red hat or a blue hat.
- For each k, P_k can see the hat of person P_l ($l > k$), but not the hat of person P_m ($m \leq k$).
- At a set time, everyone has to guess the color of her own hat by crying out "Red!" or "Blue!" Everyone is to speak simultaneously, so that nobody's guess can be informed by what others say.
- If at most finitely people answer incorrectly, everyone will be spared. Otherwise, everyone will be shot.

Problem: Identify a strategy, to be agreed upon in advance by $P_1, P_2, \ldots$, which *guarantees* that at most finitely many people answer incorrectly (and therefore that everyone is spared).

That a strategy should exist seems astonishing, since it's natural to think that choices are made in the absence of any information about the color of one's hat.

2. Strategy Setup

- Represent an assignment of hats to individuals as an ω-sequence of zeroes and ones.
- *Notation:* let S be the set of all ω-sequence of zeroes and ones; let $@ \in S$ be the ω-sequence that was actually chosen.
- Count $s, s' \in S$ as "equivalent" if there are at most finitely many k such that $s^k \neq s'^k$.
- Use this equivalence relation to partition S. (If $s \in S$, we let X_s be the equivalence class to which s belongs.)

A zero in the kth position means that P_k's hat is red; a one the kth position means that P_k's hat is blue.

$\langle 0, 0, 0, 0 \ldots \rangle$ is equivalent to $\langle 1, 0, 0, 0 \ldots \rangle$, but not to $\langle 1, 0, 1, 0 \ldots \rangle$.

3. The Strategy

- *Trick:* Let $P_1, P_2, \ldots$ agree in advance on a *representative* $\rho(X)$ from each equivalence class X of S.
- *Observation:* Even though each of $P_1, P_2, \ldots$ has limited information about $@$, the identity of $X_@$ is common knowledge.
- *Strategy:* Each person is to answer as if $\rho(X_@)$ were the actual sequence.

This requires the Axiom of Choice.

This is weird and surprising, but ultimately unproblematic.

- *Upshot:* Since $\rho(X_@)$ and @ are in the same equivalence class, they disagree only finitely. So at most finitely many people will answer incorrectly.

4. Why the existence of a strategy seems paradoxical

   - It is natural to suppose that choices are made in the absence of any information about the color of one's hat. If this is right, how they *guarantee* that at most finitely many people will answer incorrectly?
   - *A general observation in response:* following a group strategy can lead to coordination which is beneficial to the group even if individual members of the group do not increase their chances of answering correctly. This is illustrated by the following example.

B. Example: the three prisoners

I learned of this puzzle thanks to Rohit Parikh.

1. The Puzzle

   - There are three prisoners: P_1, P_2 and P_3. Each of them is wearing a hat. For each prisoner, a coin was flipped to determine whether to give her a red hat or a blue hat.
   - Each prisoner can see the colors of others' hats, but not the color of her own.
   - At a set time, everyone has to guess the color of their hat, by crying out "Red!", "Blue!", or "Pass!". The prisoners must speak simultaneously, so that nobody's guess can be informed by what others say.
   - Consequences are as follows:
     - If all three prisoners refuse to guess the color of their hat, by saying "Pass!", then everyone is killed.
     - If one of them guesses incorrectly by saying "Red!" or "Blue!", then everyone is killed.
     - If at least one prisoner guesses correctly, and nobody guesses incorrectly, then everyone is set free.

   Problem: Identify a strategy, to be agreed upon in advance, which guarantees a chance of survival above 50%.

A strategy that gives the prisoners a 50% chance of survival: select a "captain", and agree that only the captain is to offer an answer.

2. A strategy exists

   - If you see that the other two prisoners both have red hats, answer "Blue!"

- If you see that the other two prisoners both have blue hats, answer "Red!".
- If you see that the other two prisoners have hats of different colors, say "Pass!".

If all three prisoners follow this procedure, their chance of survival will be 75%. To see this, note that there are eight possible hat distributions, and that all three prisoners are set free in six of those eight possibilities (i.e. 75%):

P_1	P_2	P_3	Result
Red	Red	Red	Everyone answers incorrectly
Red	Red	Blue	P_3 answers correctly
Red	Blue	Red	P_2 answers correctly
Red	Blue	Blue	P_1 answers correctly
Blue	Red	Red	P_1 answers correctly
Blue	Red	Blue	P_2 answers correctly
Blue	Blue	Red	P_3 answers correctly
Blue	Blue	Blue	Everyone answers incorrectly

C. Comparing the two puzzles

1. The three-prisoner case

   - Suppose P_i sees two hats of the same color. Then following the agreed-upon strategy does *not* increase the probability that P_i answers correctly, over answering at random:

   $$P([\text{Hat}^i \neq \text{Hat}^j] \mid I_i) = \tfrac{1}{2} \quad (\text{for } i \neq j)$$

   I_i: P_i's information before answering

   - So: the group improves its chances of collective success not by increasing the chances of success when an answer is given, but by *coordinating* failures.

2. The ω-sequence hat case

   - Fix ρ and @. On the most reasonable assumptions I can think of, we have:

   $$P([\rho(X_@)^k \text{ accurate}] \mid I_k) = \frac{1}{2}$$
   $$P([\rho(X_@)^k \text{ accurate}]) = \frac{1}{2}$$

   where:

Both equations assume a Natural Density probability distribution; the second equation additionally assumes a reverse-binary ordering of $X_@$. See Section E for details.

  - $[\rho(X_@)^k \text{ accurate}] = \left\{s \in X_@ : s^k = \rho(X_@)^k\right\}$
    (This corresponds, intuitively, to the information that $\rho(X_@)$ accurately predicts the kth member of the actualized sequence.)
  - $I_k = \{s \in X_@ : s^n = @^n(n > k)\}$
    (This corresponds, intuitively, to the information P_k acquires when she opens her eyes.)

- So: on natural assumptions, following the strategy does *not* increase the probability of answering correctly over answering at random.

- So: the group improves its chances of collective success not by increasing the chances of individual success, but by *coordinating* their responses in the right sort of way.

3. A caveat about probabilities in the ω-sequence case

   - Our result that $P([\rho(X_@)^k \text{ accurate}]) = 1/2$ assumes a "natural" ordering of $X_@$. But in the general case, one cannot show that $X_@$ can be well-ordered without assuming the Axiom of Choice (Section F). So there's a sense in which no well-ordering of $X_@$ should really be counted as natural.

Specifically: it assumes a reverse-binary ordering, which can be defined in a natural way, *given a representative from* $X_@$.

4. An important disanalogy between the two cases

   - In the three-prisoner puzzle, looking at other people's hats does not deliver common knowledge: there is no "world" in possibility space such that it is common knowledge that that world can be ruled out.
   - The ω-sequence hat puzzle, in contrast, relies essentially on common knowledge.

D. Moral: A Diagnosis of Bacon's Puzzle

- On a reasonable way of assigning probabilities, following the strategy does *not* increase the probability of answering correctly over answering at random. In both cases, the probability is $1/2$.

- But: one shouldn't lean too heavily on probabilities here: as illustrated by caveat C(3), our probability assignments are not on very solid ground.

- Focus instead on the fact that $P_1, P_2, \ldots$ are able to acquire an interesting piece of common knowledge and use it to coordinate.

E. Calculating probabilities in the ω-sequence hat puzzle

1. Ordering equivalence classes

   - Start by considering the equivalence class containing $\langle 0,0,0,\ldots\rangle$. Its members can be ordered following reverse binary notation:

$$\begin{array}{c}\langle 0,0,0,0,\ldots\rangle\\ \langle 1,0,0,0,\ldots\rangle\\ \langle 0,1,0,0,\ldots\rangle\\ \langle 1,1,0,0,\ldots\rangle\\ \langle 0,0,1,0,\ldots\rangle\\ \langle 1,0,1,0,\ldots\rangle\\ \langle 0,1,1,0,\ldots\rangle\\ \langle 1,1,1,0,\ldots\rangle\\ \vdots\end{array}$$

   - This strategy can be generalized to an arbitrary equivalence class X, once a representative $\rho(X)$ has been selected for X. Simply use the ordering above, after applying the following transformation to each sequence in X:

$$< a_1, a_2, a_3, \ldots > \longrightarrow \left\langle |a_1 - \rho_1^X|, |a_2 - \rho_2^X|, |a_3 - \rho_3^X|, \ldots \right\rangle$$

     We call such an ordering of X the *reverse-binary ordering* (generated by ρ).

In the general case, selecting representatives requires the Axiom of Choice.

$\rho(X) = \langle \rho_1^X, \rho_2^X, \rho_3^X, \ldots\rangle$.

2. The Natural Density probability distribution for the countable set $\{s_1, s_2, s_3, \ldots\}$:

$$p(X|Y) =_{df} \lim_{n\to\infty}\left(\frac{|X \cap Y \cap \{s_1, s_2, \ldots, s_n\}|}{|Y \cap \{s_1, s_2, \ldots, s_n\}|}\right)$$

$$p(X) =_{df} p(X|\{s_1, s_2, s_3, \ldots\})$$

   (Note that this distribution makes essential use of the suggested ordering of $\{s_1, s_2, s_3, \ldots\}$.)

Two observations: (1) $p(X)$ is finitely additive but not countably additive, since $p(\{s_1, s_2, s_3, \ldots\}) = 1$ but $p(\{s_k\}) = 0$ for each k. (2) p is not well-defined for arbitrary subsets of $\{s_1, s_2, s_3, \ldots\}$. For instance, it is not well-defined fir the set of s_k such that $2^m \leq k < 2^{m+1}$, for m even.

3. We get reasonable results in the Hat Case when we use the Natural Density distribution:

   - Fix ρ and @ and define $[\rho(X_@)^k$ accurate$]$ and I_k as above. For $\langle s_1, s_2, s_3, \ldots\rangle$ an arbitrary ω-ordering of $X_@$:

Proof: only finitely many members of $X_@$ are compatible with I_k, and half of them are in $\{s \in X_@ : s^k = \rho(X_@)^k\}$.

$$P([\rho(X_@)^k \text{ accurate}] \mid I_k) = \lim_{n\to\infty}\left(\frac{\left|\left\{s \in X_@ : s^k = \rho(X_@)^k\right\} \cap I_k \cap \{s_1, \ldots, s_n\}\right|}{|I_k \cap \{s_1, \ldots, s_n\}|}\right) = 1/2$$

- In the special case where $\langle s_1, s_2, s_3, \ldots \rangle$ is a reverse-binary ordering of $X_@$, we also get:

Proof: since $\langle s_1^\tau, s_2^\tau, s_3^\tau, \ldots \rangle$ is a reverse-binary ordering of $X_@$, each cycle of $n2^k \leq i < (n+1)2^k$ ($n \in \mathbb{N}$) will be such that half of the s_i^τ take zeros as values and half take ones.

$$P([\rho(X_@)^k \text{ accurate}]) = \lim_{n\to\infty} \left(\frac{\left| \left\{ s \in X_@ : s^k = \rho(X_@)^k \right\} \cap \{s_1^\tau, \ldots, s_n^\tau\} \right|}{|\{s_1^\tau, \ldots, s_n^\tau\}|} \right) = 1/2$$

- In the general case, however, $P([\rho(X_@)^k \text{ accurate}])$ can take arbitrary values in $[0, 1]$.

  Note that when $P([\rho(X_@)^k \text{ accurate}])$ has a value other than $\frac{1}{2}$ (or is undefined), we get gross failures of conglomerability. Let I_k^i ($i \in \mathbb{N}$) be the family of (2^k-membered) subsets of $X_@$ such that for some $s' \in X_@$:

To get probability $1/n$, choose $\langle s_1, s_2, s_3, \ldots \rangle$ such that $(s_i)^k = \rho(X_@)^k$ iff i is divisible by n.

$$I_k^i = \{s \in X_@ : s(n) = s'(n) \ (n > k)\}$$

  Since the I_k^i form a partition of $X_@$, the following constitutes a failure of conglomerability:

$$P([\rho(X_@)^k \text{ accurate}] \,|I_k^i) = \frac{1}{2} \ (i \in \mathbb{N}) \qquad P([\rho(X_@)^k \text{ accurate}]) \neq \frac{1}{2}$$

  And, as noted earlier, the first of these equations is independent of one's ω-ordering of $X_@$.

F. Orderings of equivalence classes cannot be explicitly characterized

We show that it is impossible to explicitly characterize a relation $<$ such that each equivalence class X of S is well-ordered by $<$.

Suppose, for *reductio*, that one can explicitly characterize $<$, and therefore that one can characterize a choice set for S without using the Axiom of Choice. We proceed by verifying that such a choice set can be used to characterize a set that is not Lebesgue measurable. This suffices the prove our result because Solovey showed that one can only prove the existence of sets that are not Lebesgue measurable using the Axiom of Choice.

The construction is analogous to that of the Vitali Theorem. It is easy to verify that there is a bijection between $S^\star$ and the real numbers in $[0, 1)$: simply assign each $s \in S^\star$ to the real number "$0.s(0), s(1), \ldots$". I will henceforth speak interchangeably of real numbers and the corresponding members of $S^\star$.

$S^\star$ is the result of removing from S sequences equivalent to $\langle 1, 1, 1, \ldots \rangle$. There is no loss of generality because if a choice set exists our partition of S, it must also exist for our partition of $S^\star$.) I Because we are working with $S^\star$ rather than S, we don't have to worry about numbers with more than one binary expansion.

Let X be an equivalence class and let $a, b \in X$. Assume with no loss of generality that $a \leq b$. We can now make two observations:

1. a and b are both real numbers, so $b - a$ must also be a real number. Since a and b's binary expansions match after a certain point, the binary expansion of $b - a$ must terminate, so $b - a$ must be a rational number.

2. If q is a rational number whose binary expansion terminates, then $a + q$ (or $(a + q) - 1$, if $a + q \geq 1$) is a member of X.

These two observations allow us to adapt the proof that the Vitali sets are non-measurable to show that a choice set C for our partition of $S^\star$ must be non-measurable.

Briefly, for each rational number q whose binary expansion terminates, the set C^q:

$$\left\{ x : \exists c \in C, x = \begin{cases} c + q & (c + q < 1) \\ (c + q) - 1 & (c + q \geq 1) \end{cases} \right\}$$

will be a translation of C around the unary circle. So it follows from the uniformity condition on Lebesgue measure that C and C^q must have the same Lebesgue measure, if they have Lebesgue measure at all. But $S^\star$ is partitioned into countably many C^q (q terminating), so it follows from Countable Additivity that there is no single Lebesgue measure all C^q could have.

References

Andrew Bacon. A paradox for supertask decision makers. *Philosophical Studies*, 153(2):307–311, 2011.