# Doomsday and the Dice Room Murders

Kevin S. Van Horn

July 28, 2015

## 1 The puzzle

Scott Aaronson has a wonderful book called *Quantum Computing Since Democritus* [1], and in his chapter on the anthropic principle he writes about a thought experiment (attributed to John Leslie) that he calls the *Dice Room*:

> Imagine that there's a very, very large population of people in the world, and that there's a madman. What this madman does is, he kidnaps ten people and puts them in a room. He then throws a pair of dice. If the dice land snake-eyes (two ones), then he simply murders everyone in the room. If the dice do not land snake-eyes, then he releases everyone, then kidnaps 100 people. He now does the same thing: he rolls two dice; if they land snake-eyes, then he kills everyone, and if they don't land snake-eyes, then he releases them and kidnaps 1000 people. He keeps doing this until he gets snake-eyes, at which point he's done.

Think of the population of this world as being randomly partitioned into a sequence of batches, with batch $t$ having $10^t$ people. On step $t$ the madman kidnaps the people in batch $t$.

The Dice Room is a metaphor for human extinction. The population of this hypothetical world corresponds to all human beings who ever have or who ever may live. The batches represent succeeding generations of humanity, the exponentially growing batch size represents exponential growth of the human population, the madman represents the sum of the various existential risks facing humanity, and rolling snake-eyes corresponds to an extinction event that wipes out humanity. Being kidnapped and placed in the Dice Room corresponds to being born into a particular generation of humanity.

This Dice Room scenario leads to the following question: if you are kidnapped, what is your probability of dying? Correspondingly, how likely is ours to be the generation in which humans go extinct? Aaronson gives two answers. The first:

> ...the dice have a 1/36 chance of landing snake-eyes, so you should only be a "little bit" worried (considering.)

Let's call this the "proximate risk" argument.

And the second:

> ...consider, of people who enter the room, what the fraction is of people who ever get out. Let's say that it ends at 1000. Then, 110 people get out and 1000 die. If it ends at 10,000, then 1110 people get out and 10,000 die. In either case, about 8/9 of the people who ever go into the room die.
>
> ... We can say that we're conditioning on a specific termination point, but that no matter what that point is, we get the same answer. It could be 10 steps or 50 steps, but no matter what the termination point is, almost all of the people who go into the room are going to die, because the number of people is increasing exponentially.

Let's call this the "proportion murdered" argument. Also, $1000/1110 \approx 9/10$ rather than 8/9, so we'll use 9/10 in the rest of this note. The proportion-murdered argument amounts to an argument that human extinction is probably nigh.

Aaronson comments that "If you're a Bayesian, then this kind of seems like a problem," as Bayesian reasoning seems to give two different answers to the same question given the same information.

## 2 In a nutshell: how imminent doom is averted

Right off the bat, you should be smelling something rotten here: Bayesian reasoning is just unadorned probability theory, nothing more and nothing less, and we know that the rules of probability theory are logically consistent—or rather, if they're not, then Zermelo-Fraenkel set theory is also inconsistent and mathematicians start throwing themselves off of tall buildings. If we have two different arguments giving different answers, one of them is wrong. In particular, the proportion-murdered argument is wrong.

The proportion-murdered argument relies on the property that $\pi_t$, the prior probability that you are in batch $t$, increases exponentially with $t$. If the madman murders at step $T$ and $\pi_{t+1} = 10\pi_t$ for all $1 \leq t < T$, and you know you are kidnapped at some point, then the probability that you are in batch $T$ (and hence die) is

$$\frac{\pi_T}{\sum_{t=1}^{T} \pi_t} \approx \frac{9}{10}.$$

But such exponential growth cannot continue indefinitely: the probabilities $\pi_t$ must sum to 1, and *any infinite sequence of nonnegative numbers that sums to 1 must go to zero in the limit.* Regardless of what prior distribution we use for your batch number, we know that $\pi_t \to 0$ as $t \to \infty$. Thus is the proportion-murdered argument destroyed, imminent extinction averted, and the foundations of mathematics rescued.

In the rest of this note I'll flesh out the analysis.

## 3 Let's get explicit

To start the analysis, let's be crystal clear on our assumptions:

1. Writing $M(t)$ for "the madman murders at step $t$" and $\mathrm{hm}(t)$ for "the madman has already murdered by step $t$," the madman's choice process is as follows, for $t \geq 1$:

$$\begin{aligned} \mathrm{hm}\,(1) &= \text{false} \\ \mathrm{hm}\,(t+1) &= (\mathrm{hm}\,(t) \ \text{ or } M\,(t)) \\ \Pr\left(M\,(t) \mid \mathrm{hm}\,(t)\right) &= 0 \\ \Pr\left(M\,(t) \mid \neg\mathrm{hm}\,(t)\right) &= p \\ &= \frac{1}{36} \end{aligned}$$

   where "¬" means "not." That is, once the madman has murdered he does not do so again, and if he has not murdered in a prior step then there is a probability $p$ he will murder in the current step.

2. The population is partitioned into batches, which are numbered from 1 onward, with every individual belonging to exactly one batch. We write $b$ for the specific batch to which you belong.

3. We write $\pi_t$ for $\Pr(b = t)$, the prior probability that you belong to batch $t$.

4. You die, written $D$, if the madman murders on the step corresponding to your batch number:

$$D \equiv M\,(b)\,.$$

5. You are kidnapped, written $K$, if the madman has not yet murdered on the step corresponding to your batch number:

$$K \equiv \neg \mathrm{hm}\,(b)\,.$$

6. You know that you are kidnapped, but not your batch number. Thus, the probability of interest is

$$\Pr\,(D \mid K)\,.$$

What is the specific set of batch probabilities $\pi_t$ that arise in the Dice Room scenario? We could derive it as follows:

- Assume that batch $t$ contains $10^t$ individuals and that there are $n$ batches (a finite number). This gives a total population size of $N = \sum_{t=1}^{n} 10^t$.
- Number the individuals from 1 to $N$; batch 1 consists of the first $10^1$ individuals, batch 2 consists of the next $10^2$ individuals, etc.
- Let $i$ be your assigned number; since you don't know your number, the prior for $i$ is the uniform distribution over the integers from 1 to $N$.

This gives $\pi_t = 10^t/N$ for $1 \leq t \leq n$ and $\pi_t = 0$ for $t > n$. In this case if the madman "murders" at a step $t > n$ then nobody actually dies, as all batches after batch $n$ are empty.

What if we don't want to assume a finite population? Then things get trickier, because *there does not exist a uniform distribution over the infinite set of all positive integers.* We would have to give some prior on $i$ that was at least weakly informative.

The details don't matter. The analysis that follows doesn't depend on what specific batch probabilities $\pi_t$ are used, as long as they are valid probabilities: each is nonnegative and they sum to 1.

## 4 The proximate-risk argument

The first step in formalizing the proximate-risk argument is to decompose the problem by the batch to which you belong:

$$\Pr\,(D \mid K) \quad = \quad \Pr\,(M\,(b) \mid K)$$

$$
\begin{aligned}
&= \sum_{t=1}^{\infty} \Pr\left((b=t)\ \&\, M(t) \mid K\right) \\
&= \sum_{t=1}^{\infty} \Pr\left(b=t \mid K\right) \cdot \Pr\left(M(t) \mid (b=t)\ \&\, K\right)
\end{aligned}
$$

*In words:* take the weighted average, over all $t$, of the probability that the madman murders at step $t$ if you are in batch $t$ and are kidnapped; use as the weighting the probability that you are in batch $t$ if you are kidnapped.

Now

$$
\begin{aligned}
(b=t)\ \&\, K &\Leftrightarrow (b=t)\ \&\, \neg \mathrm{hm}(b) \\
&\Leftrightarrow (b=t)\ \&\, \neg \mathrm{hm}(t)
\end{aligned}
$$

so

$$
\begin{aligned}
\Pr\left(M(t) \mid (b=t)\ \&\, K\right) &= \Pr\left(M(t) \mid (b=t)\ \&\, \neg \mathrm{hm}(t)\right) \\
&= \Pr\left(M(t) \mid \neg \mathrm{hm}(t)\right) \\
&= \frac{1}{36}
\end{aligned}
$$

with the second step justified by the fact that $(b=t)$ and $M(t)$ are independent propositions, even when conditioning on $\neg \mathrm{hm}(t)$; then

$$
\begin{aligned}
\Pr\left(D \mid K\right) &= \sum_{t=1}^{\infty} \Pr\left(b=t \mid K\right) \cdot \frac{1}{36} \\
&= \frac{1}{36} \Pr\left(\exists t \geq 1.\, b=t\right) \\
&= \frac{1}{36}.
\end{aligned}
$$

So the proximate-risk argument checks out as valid.

## 5 The proportion-murdered argument

Formalizing this argument proceeds in the same fashion as for the proximate-risk argument, except that we decompose the problem by the time step at which the madman murders:

$$
\begin{aligned}
\Pr\left(D \mid K\right) &= \Pr\left(M(b) \mid K\right) \\
&= \sum_{t=1}^{\infty} \Pr\left((b=t)\ \&\, M(t) \mid K\right) \\
&= \sum_{t=1}^{\infty} \Pr\left(M(t) \mid K\right) \cdot \Pr\left(b=t \mid M(t)\,\&\, K\right)
\end{aligned}
$$

*In words:* take the weighted average, over all $t$, of the probability that you are in batch $t$ if you are kidnapped and the madman murders at step $t$; use as the weighting the probability that the madman murders at step $t$ given that you are kidnapped.

Let's verify that this really is a weighted average, i.e., that the weights sum to 1. We have

$$\Pr(M(t)) = p(1-p)^{t-1}$$

and hence

$$\begin{aligned}\Pr(\exists t \geq 1.\, M(t)) &= \sum_{t=1}^{\infty} \Pr(M(t)) \\ &= p \sum_{t=1}^{\infty} (1-p)^{t-1} \\ &= 1.\end{aligned}$$

It is true in general that if $\Pr(Y) = 1$ then $\Pr(Y \mid Z) = 1$ for any $Z$ whose probability is not 0. So

$$\begin{aligned}\sum_{t=1}^{\infty} \Pr(M(t) \mid K) &= \Pr(\exists t \geq 1.\, M(t) \mid K) \\ &= 1.\end{aligned}$$

*In words:* the madman is guaranteed to eventually murder, and conditioning on the fact that you are kidnapped does not change this.

Note that

$$M(t) \,\&\, K \quad \Leftrightarrow \quad M(t) \,\&\, (b \leq t);$$

if the madman murders at step $t$, you are kidnapped if and only if your batch number is $t$ or less. Defining

$$Q_t \equiv \frac{\pi_t}{\sum_{u=1}^{t} \pi_u},$$

we then find that

$$\begin{aligned}\Pr(b = t \mid M(t) \,\&\, K) &= \Pr(b = t \mid M(t) \,\&\, (b \leq t)) \\ &= \Pr(b = t \mid b \leq t) \\ &= Q_t\end{aligned}$$

(This uses the fact that $b = t$ and $M(t)$ are independent propositions even if we condition on $(b \le t)$.)

*In words:* if the madman murders at step $t$ and you are kidnapped, then $Q_t$ is the probability that you are in batch $t$, exactly as assumed by the proportion-murdered argument.

We then have

$$\Pr(D \mid K) = \sum_{t=1}^{\infty} \Pr(M(t) \mid K) \cdot Q_t;$$

that is, the probability we seek to determine is the weighted average of $Q_t$ over all $t$.

So far every step of the proportion-murdered argument has checked out. However, the proportion-murdered argument also relies on a claim that $Q_t \approx 9/10$ regardless of $t$. That is, if it is known that you are in one of the first $t$ batches, it is highly likely that you are in batch $t$ itself. If $\pi_{u+1} = 10\pi_u$ for $1 \le u < t$, then $Q_t$ does indeed work out to approximately 9/10. But, as previously noted, it cannot be true that $\pi_{u+1} = 10\pi_u$ for all $u \ge 1$: the probabilities $\pi_u$ can sum to 1 only if $\pi_u \to 0$ as $u \to 0$. From this we see that $Q_t \to 0$ as $t \to \infty$.

## 6 Detail: fixing the proportion-murdered argument

Let's carry out the rest of the analysis to verify that, even taking the proportion-murdered approach to the question, we still get an answer of 1/36.

First, some results we will need:

$$\begin{aligned}
\Pr(K) &= \Pr(\neg\mathrm{hm}(b)) \\
&= \sum_{t=1}^{\infty} \Pr((b = t)\ \&\ \neg\mathrm{hm}(t)) \\
&= \sum_{t=1}^{\infty} \Pr(b = t) \Pr(\neg\mathrm{hm}(t)) \\
&= \sum_{t=1}^{\infty} \pi_t (1-p)^{t-1}
\end{aligned}$$

and

$$\Pr(M(t)\,\&\,K) = \Pr(M(t)\,\&\ (b \le t))$$

$$
\begin{aligned}
&= \Pr\left(M(t)\right) \cdot \Pr\left(b \le t\right) \\
&= p\left(1-p\right)^{t-1} \sum_{u=1}^{t} \pi_u .
\end{aligned}
$$

Then

$$
\begin{aligned}
\Pr\left(M(t) \mid K\right) \cdot Q_t &= \frac{\Pr\left(M(t) \,\&\, K\right) \cdot Q_t}{\Pr\left(K\right)} \\
&= \frac{p\left(1-p\right)^{t-1} \pi_t}{\Pr\left(K\right)}
\end{aligned}
$$

and so

$$
\begin{aligned}
\Pr\left(D \mid K\right) &= \sum_{t=1}^{\infty} \Pr\left(M(t) \mid K\right) \cdot Q_t \\
&= \sum_{t=1}^{\infty} \frac{p\left(1-p\right)^{t-1} \pi_t}{\Pr\left(K\right)} \\
&= p \frac{\Pr\left(K\right)}{\Pr\left(K\right)} \\
&= p \\
&= \frac{1}{36}.
\end{aligned}
$$

## 7 Concluding comments

The take-away here is this: don't trust purely verbal arguments when it comes to determining conditional probabilities. It is too easy to make mistakes, even for experts.

Experienced software developers take it for granted that any piece of code that hasn't been tested is wrong. A similar attitude is warranted towards any argument involving conditional probabilities. The counterpart to testing code is to formalize the argument. In other words, *do the math*, justifying each step of the derivation by appeal to the laws of probability.

## References

[1] Aaronson, Scott (2013). *Quantum Computing Since Democritus*. Cambridge University Press.