THE DOOMSDAY SIMULATION ARGUMENT.

OR WHY ISN'T THE END NIGH AND YOU ARE NOT LIVING IN A SIMULATION.

BY

ISTVÁN A. ARANYOSI
Philosophy, Central European University
E-mail: *fphari01@phd.ceu.hu*

There has been a revigoration in recent years of discussions surrounding two independent traditional philosophical topics, but from a new perspective, and a shared analytical ground. The two topics are: the end of the world, and scepticism about the external world; the common analytical ground is that of anthropic reasoning. More precisely, the topics I have in mind regard the probability we should assign to our species going extinct in a short time to come, and the probability we should assign to the hypothesis that we are living in a Matrix-like computer simulation. As regards anthropic reasoning, it is customary to understand it as the formulation of an anthropic principle; we will adopt a formulation due to John Leslie (1996, p. 190): "observers can most expect to find themselves in the spatiotemporal regions where most of them are found".

In this paper I will propose and discuss an argument (1) combining the ideas typical to two previously exposed arguments —the Doomsday Argument (DA) and the Simulation Argument (SA)—but (2) having, at the same time, a conclusion contrary to both those of the aforementioned arguments, which I will call 'the Doomsday Simulation Argument (DSA)'. I will first briefly

expose DA and SA, then I will formulate and defend the premises and analyse the conclusions of DSA.

DOOMSDAY

Let us then start with DA (Carter 1983, Leslie 1996). It is a fact that the human population is growing. By the end of 2004, there will be about 8 billion humans on Earth. According to some estimations, by 2010 there will be more than 10 billion humans[1], while in 2100 the human population will reach 27 billions (Leslie 1996); as for the past, a realistic estimate of the number of humans that have ever lived so far is that of approximately 90 billion. Thus our population in 2004 represents almost 10 per cent of all the humans that have lived so far. Applying the anthropic principle to one's position in time, we should judge as more probable that we are in a temporal niche where most of the observers are located. More exactly, if we suppose that we are among the earliest 10 per cent of all the humans that will ever have lived, then the number of all the humans that will ever have lived (past, present, and future) is 900 billion. This number may be reached by 2500, which means that in this case we should expect the extinction of the human species in a few centuries. If, on the other hand, we expect to survive for millions of more years, then we are among less than 0.00000001 of all humans. According to the anthropic principle we should consider as more probable that we are among the 0.1 than among less than 0.00000001 of all humans. Hence, we should assign a proportionally higher probability to DOOM SOON!

For more clarity, let us formulate first the anthropic principle that plays the main role in DA, which we will call the Strong Indifference Principle (SIP):

If we knew that a fraction x of all observers who will have ever lived are among those that have ever lived until and are alive at the time we make the observation, then, even if we know that we are in 2004, our credence that we are among the earliest x of all observers should be equal to x.

SIP says, in effect, that one should consider as if one were a random sample from the set of all people who will ever have lived. In other words, one

[1] Levin (1996, 570), cited by Smith (1997)

should reason as if being indifferent with respect to where one's observed birth rank is located in time within the totality of birth ranks of all humans who will ever have lived.

We call it strong because of the clause I have emphasised in the above formulation. And this is one reason for doubting the soundness of DA, which I will shortly continue to expose. The reason for this doubt is the limited reasonableness of such a clause: given that we know our temporal location, the requirement of reasoning as if we were a random sample of the set of all people, past, present, and future, is harder to accept.

The next step in the argument is the application of Bayes's theorem for belief updating. Bayes's theorem relates four probabilities, given a hypothesis H and evidence E: the prior probability of H, p(H); the probability of E given H, p(E|H); the probability of E given nonH, p(E|nonH); and the posterior (the updated) probability of H given E, p(H|E). The formula is:

$$p(H \mid E) = \frac{p(H)p(E \mid H)}{p(H)p(E \mid H) + p(nonH)p(E \mid nonH)}$$

Let us then assign interpretations to H and E, and values to the probabilities. Our hypothesis H is that the human species will go extinct by, say, 2200, while E is the proposition that one is alive in 2004. Further, the negation of H will be considered equivalent to the truth of the proposition that we will survive for thousands of centuries, and H|E will represent the truth of H conditional on the truth E. Suppose our credence that the human species will go extinct by 2200 is only 1 per cent. Then we have the following assignment of values to the variables:

p(H) = 0.01
p(nonH) = 0.99
p(E|H) = 0.1
p(E|nonH) = 0.001

Replacing them in Bayes's formula we obtain:

$$p(H \mid E) = \frac{0.01 \times 0.1}{0.01 \times 0.1 + 0.99 \times 0.001} \approx 0.503$$

which means that we should update our prior from 1% to slightly more than 50% chance of DOOM BY 2200! We can further observe that the more time we expect to survive, the higher will be the posterior probability of H given E. For example, if we expect to be around for millions of more centuries, the value of p(H|E) will approach unity.

SIMULATION

SA has been proposed by Nick Bostrom (2003), and is based on two assumptions. The first is that of substrate-independence of consciousness, which means that consciousness supervenes on any of a broad range of physical realizations, provided they implement the right sort of computational structures and processes. In other words, if one has the resources to implement sufficiently complex computational structures and processes, one is able to simulate consciousness[2]. The second assumption is that if our technological progress will continue for a sufficiently long time with the same pace as so far, then humankind will attain a posthuman stage of civilisation --with a maximal level of technological capabilities that one can currently consider as consistent with all the physical laws and all the material and energy constraints of our universe—and will be able, due to an immense computing power, to simulate a huge number of entire ancestor civilisations, including the mental processes that are manifested within them.

Given these assumptions, the basic idea of SA is expressible by the following question, as Bostrom puts it: "if there were a substantial chance that our civilization will ever get to the posthuman stage and run many ancestor-simulations, then how come you are not living in such a simulation?"

More formally, let us use the following notation:

[2] Of course, one need not understand substrate-independence as necessary; it may be a contingent fact about our laws of nature, and so it is compatible with mind-body dualism.

f_p = Fraction of all human-level technological civilizations that survive to reach a posthuman stage

$\overline{H}$ = Average number of individuals that have lived in a civilization before its reaching a posthuman stage

$\overline{F}$ = Average number of ancestor-simulations run by a posthuman civilization

The actual fraction of all observers with human-type experiences that live in simulations is

$$f_{sim} = \frac{f_p \overline{H}\,\overline{N}}{f_p \overline{H}\,\overline{N} + \overline{H}}$$

and dividing by $\overline{H}$ we get :

$$f_{sim} = \frac{f_p \overline{N}}{f_p \overline{N} + 1}$$

Denoting the fraction of posthuman civilizations that contain at least some individuals who are interested in running ancestor-simulations and have the resources to run a significant number of these by f_I, and the average number of ancestor-simulations run by such civilisations by $\overline{N_I}$, we have

$$\overline{N} = \overline{f_I}\,\overline{N_I}$$

which gives us:

$$f_{sim} = \frac{f_p \overline{f_I}\,\overline{N_I}}{f_p \overline{f_I}\,\overline{N_I} + 1}$$

Since, given the immense computing power of future posthuman civilisations, $\overline{N_I}$ is extremely large, f_{sim} will approach unity, unless f_p, f_I, or both approach zero. This means that at least one of the following three propositions must be true:

(1) It is very likely that we go extinct before reaching a posthuman stage ($f_p \approx 0$).
(2) It is very unlikely that some posthuman civilisation will contain at leat some individuals who are interested in and have resources to run a significant number of ancestor-simulations ($f_I \approx 0$).
(3) It is almost sure that we are living in a computer simulation ($f_{sim} \approx 1$).

The principle that permits the assertion of (3) is similar to SIP, but also different from it in one respect; we will call it the Weak Indifference Principle (WIP).

Weak Indifference Principle (WIP):

If we knew that a fraction x of all observers with human-type experiences live in simulations, then, without knowing whether our own experiences are more likely to be biologically implemented than artificially simulated, our credence that we are living in a simulation should be equal to x.

WIP says, similarly to SIP, that one should consider as if one were a random sample from the set of all people whith human-type experiences. In other words, one should reason as if being indifferent with respect to where one's observed civilisation is located within the cyber-cum-natural space of all civilisations.

The reason why we call it 'weak', and the difference between it and SIP, lies, again, in the emphasised clause, which states the indistinguishability from our perspective as observers of real from simulated experiences.

DOOMSDAY SIMULATION

As I mentioned above, SIP states the implausible requirement that we should neglect the known fact that we actually are in 2004, and reason as if we were a random sample from the set of all the people who wil ever have have lived. The problem I see with WIP is that whilst it is certainly weaker, in the sense I explained, than SIP, there is a sense in which it is stronger than it. More exactly, in order for it to be usable in SA we have to assume not only that we don't know which population we belong to (one of the many simulated ones or one of the few real ones), given the indistinguishability of simulated and real minds, but also, even if implicitly, that we don't know which *time* we are actually living. SA shares with DA one important factor or variable: time. In DA we know the time slice we are ocuppying, and nevertheless prescribe indifference with respect to our temporal location. In SA time is also important; reaching a posthuman stage requires time, and so the more time we expect to survive, the more probable that a huge number of simulations are actually run. To accept WIP as a premise for a valid SA means, in effect, to accept that we don't know that we are in 2004. But this should be, as far as I see, an assumption of ours for SA to get off the ground; the only thing we should have no information about is whether we are simulated or real, not whether we are at the beginning of the 21st century or not.

The problems for DA and SA can be summarized as follows: DA is based on SIP, which is too strong to be asserted, and SA is based on WIP, which is sufficiently weak to be asserted, but requiring a too strong further assumption to be conjoined with in order for SA to go through. SIP can be represented as:

$(\mathrm{K}p \,\&\, \mathrm{K}q) \supset \mathrm{IND}\ \{H_1, H_2, \ldots, H_n\}$,

while WIP as

$(\mathrm{K}r \,\&\, \sim\mathrm{K}s) \supset \mathrm{IND}\ \{H^*_1, H^*_2, \ldots, H^*_n\}$,

where K is the knowledge operator, p and q, like r and s, are propositions we are in certain epistemic relations with according to the premises of DA and SA, respectively, $\{H_1, H_2, \ldots, H_n\}$ is the hypothesis space, and IND

{…} means indifference with respect to the elements of the hypothesis space {…}. The problem with SIP is the questionableness of the conditional itself, because of the K*q* component. As opposed to SIP, the conditional expressed by WIP is unproblematic, but the second conjunct of its antecedent, K*r* & ~K*s*, which should be asserted as a premise if SA is to go through, is problematic for the reasons I have given above.

In order to avoid these problems one should find a set of assumptions such that they would require a more reasonable indifference principle, which I will call the Middling Indifference Principle (MIP), having the following form:

(K*t* & K*u* & ~K*v*) ⊃ IND {H_1, H_2, …, H_n},

and the following two properties: the conditional's truth is not questionable and the conjuncts of the premise that needs to be asserted for the argument based on the principle to go through — K*t* & K*u* & ~K*v* — are unproblematic. This set of assumptions is provided by the argument I propose: the Doomsday Simulation Argument (DSA). It combines assumptions from DA and SA.

The basic assumption inherited from DA is that we know which time we are living, and that inherited from SA is that we don't know which population we belong to. These basic assumptions correspond to K*u* and K*v* in MIP; K*t* represents a supposition of DSA, just as DA's supposition is that we are among the earliest faction *x* of all people who will ever have lived, and SA's supposition is that there is a fraction *x* of observers who live in simulations. DSA's supposition is a compound proposition, based on the following idea. If a posthuman civilisation contains some individuals interested and having resources to run a huge number of simulations, then it is more probable that these simulations are run with the purpose of evaluating the risks of extinction of the simulating species and the ways to avoid it by using the simulated species as the experimental subject of exposure to such unique risks, than with the purpose of entartainment. The uniqueness of these risks is well pointed out by Bostrom (2002), who calls them 'existential risks' (and offers a very interesting classification and analysis of them):

> Our approach to existential risks cannot be one of trial-and-error. There is no opportunity to learn from errors. The reactive approach – see what happens, limit damages, and learn from experience – is unworkable. Rather, we must take a proactive approach. This requires *foresight* to anticipate new types of threats and a willingness to take decisive *preventive action* and to bear the costs (moral and economic) of such actions.

And this uniqueness creates a strong incentive for posthuman civilisations to expose *simulated* populations to such risks[3].

Further, such a simulated world will have its human population gone extinct at a certain time. We should distinguish between the simulated time and the time of the simulation; the simulation of a whole human history of, say, 100 million years may take no more than one second for the posthuman civilisation to run, given its immense computing power. This one second interval belongs to the time of the simulating society, it is the time of the simulation, while the 100 million years interval belongs to the time of the simulated world, it is the simulated time.

We shall say that a whole simulated world is 'a matrix'[4], a name inspired by the movie *The Matrix*[5], and we will call 'the order of the matrix' the

[3] There is an obvious objection that can be formulated, which my phrasing apparently invites: presumably, in a posthuman civilisation it will be considered quite unethical to experiment even with simulated people (Aranyosi 2004) and so it is improbable that such a civilisation would actually run such simulations. My reply is that by simulating an ancestor evolutionary history *without deliberately programming existential risks* to be faced by the simulated civilisation, but by letting that world to evolve in a risk environment close to the objective risk environment of the simulating civilisation, is far from the admittedly reprehensible practice of exposing simulated people to abnormally high levels of stress caused by abnormally high risks. The only difference between the simulating and the simulated civilisations is that the latter can observe and analyse the evolution of millions of generations in no more than one second, just as our present-day simulations of millions of generations of (comparatively very primitive) artificial societies, like Epstein and Axtell's (1996) SUGARSCAPE, takes a few hours. But even this advantage is a relative one: simulated civilisations may last for a long enough time to become posthuman and run their own ancestor-simulations on powerful simulated computers, "virtual machines"; the seeds of such computers are, as Bostrom (2003) points out, found in today's virtual machines, like Java script web-applets, which run on a virtual machine inside your desktop.

[4] See Chalmers (2003), for a discussion of skepticism *versus* matrix scenarios.

[5] I should note that, unlike in *The Matrix*, here we have a simulation of all the physical reality, including brains, which reality will automatically contain, given the assumption of substrate-independence, our minds.

simulated time at which the matrix's human observer population goes extinct. A matrix of order *n* will be denoted by 'matrix$_n$' and will be understood as a matrix whose human population lasts for *n* years.

Then the two basic assumptions inherited by DSA from DA and SA, respectively are: (a) that we know our simulated time we are living, and (b) that we don't know which matrix we belong to. In this way DSA, unlike its ancestors, DA and SA, can accommodate two plausible assumptions. Finally, DSA's supposition is:

We are among the earliest fraction *x*, *in some matrix*, of the totality of people that will ever have lived in that matrix, and there is a fraction *y* of all observers, who live in matrices.

For simplicity, we will asume (1) that time is discrete, with a granularity of 1 year, (2) that the huge number of matrices are uniformly distributed on the line of natural numbers, with a one-one correspondence between the numbers and the orders of matrices, and (3) that the posthuman civilisation starts all the matrices simultaneously.

Then, assumption (a) means that we know we are in a matrix of order greater or equal to 2004, and assumption (b) means that we don't know the order of the matrix we are living in. Let us then assign interpretations and values for the variables occurring in Bayes's theorem, supposing we give a probability of only 1 per cent to DOOM SOON, and start, as we did in DA, with the assumption of our being among the earliest 10 per cent in some matrix.

H = doomsday by 2200 = we are in matrix$_{n \approx 2200}$

non-H = survival for thousands of centuries = we are in matrix$_{n > k >> 2200}$

E = one is alive in 2004 = one is in matrix$_{n \geq 2004}$

E|H = one is alive in 2004, given that he is in matrix$_{n \approx 2200}$

p(H) = 0.01
p(non-H) = 0.99

p(E|H) = 0.1

The most important part to the argument is the value we should assign to the probability of being an observer in 2004 conditional on the human species surviving for thousands of centuries, p(E|non-H). Since non-H means that we are in some matrix of an order situated between the values of *n* and *k*, and since according to our MIP, SIP is an intra-matrix applicable rule, while WIP is an inter-matrix applicable one, p(E|non-H) will be obtained by summing over the fractions observers occupy in the intra-matrix sets of all observers corresponding to each matrix order between *n* and *k*:

$$p(E \mid nonH) = \sum_{i=k}^{n} f_i$$

where f_i = the fraction of all the humans who will ever have lived within which our current population is located (i.e. the fraction all humans that have ever lived up to 2004 represent in the total number of humans, past, present, and future), if we are in a matrix of order *i* (i.e. a matrix whose human population lasts for *i* years).

Applying Bayes's theorem then we have

$$p(H \mid E) = \frac{0.1 \times 0.01}{0.1 \times 0.01 + 0.99 \sum_{i=k}^{n} f_i}$$

Since

$$\lim_{(n-k)\to\infty} \sum_{i=k}^{n} f_i = 1,$$

if, $n - k >> 1$, then

$$\sum_{i=k}^{n} f_i \approx 1$$

and thus $p(H|E) \approx 0.001$. This means nothing else than that we should update our prior for DOOM SOON from 1% to 0.01 %!

For the sake of completeness, here are some calculations (assuming linear[6] population growth):

For $k - 2004 = 100{,}000$ years, $p(E|H) = p(E|\text{non-H}) = 0.1$ at $n \approx 1{,}600{,}000$ $\Leftrightarrow$ H's posterior equals its prior.
For $k - 2004 = 100{,}000$ years, $0.1 = p(E|H) << p(E|\text{non-H}) \approx 1$, at $n \in (10^7, 10^8)$ $\Leftrightarrow$ H's posterior much lower than its prior.

As a consequence, if the reasoning behind DA and SA is impeccable, then that behind DSA should be impeccable as well. At the same time, DSA fares better than its ancestors because it is based on more plausible assumptions, which make applicable a more reasonable indifference principle, MIP. Finally, and most importantly, there is nothing paradoxical about the predictions of DSA.

CONCLUSIONS OF DSA

By inspecting our last three formulae one can draw two main conclusions, which, we can further observe, buttress each other.

First conclusion: the more time we expect to be around, the less the probability we should assign for DOOM SOON, after Bayesian updating, because the more the probability of our being in a matrix of very high rather than very low order!

Second conclusion: the more time we expect to be around, the higher the probability, after many iteration of Bayesian updating, of our being in the SUPERMATRIX (the matrix of highest order = REALITY), i.e. the more

[6] Of course, if population grows exponentially, *n* will take much higher values. It is not essential, however, for the argument to assume linear growth; it will still be true on the assumption of exponential growth that there will be a value for *n* sufficiently high to render $p(E|H) << p(E|\text{non-H})$, and thus $p(H|E)$ less than $p(H)$.

probable that *it is we ourselves* (or, more exactly, those whose flesh-and-blood ancestors are *us*) who will create all or many of the matrices!

AN OBJECTION AND A REPLY

The second of our conclusions is the more intriguing. Recall my objection to SA that it has to assert the hypothesis of our lacking information about our location in time, and that this makes it less convincing. At the same time, observe that SA is consistent from the point of view of betting odds based rationality. As Bostrom points out:

> If betting odds provide some guidance to rational belief, it may also be worth to ponder that if everybody were to place a bet on whether they are in a simulation or not, then if people use the bland principle of indifference [NB, our WIP], and consequently place their money on being in a simulation if they know that that's where almost all people are, then almost everyone will win their bets. If they bet on *not* being in a simulation, then almost everyone will lose.

Now one may object to my argument that it is apparently inconsistent with betting odds based considerations. Our second conclusion prescribes my betting on not being in a simulation, but I proceeded from the assumption I share with Bostrom that there will be a huge number of simulations in the future, and so I have to agree that most of the people live in simulations. Therefore, if I bet on not being in a simulation, and I recommend it to everybody, then almost everyone will lose their bets if they follow my advice.

In response, I should first point out that the second conclusion does not recommend betting on not being in a simulation given that one knows that that's where most of the people are found, but betting on not being in a simulation given that one expects the human race to last for an extremely long time. Second, and more to the point, consider the following betting prospect:

Get (a) *$ sum equal to your bet times the number of the years you expect the species to survive if not simulated, otherwise get* (b) *$ sum equal to your bet*

times the difference in number of years between the longest lasting and your actual matrix.

I think this betting prospect is a correct representation of the situation, given our previously explained assumptions. Now consider the two hypotheses: everyone bets on not being in a simulation and everyone bets on being in a simulation. It is I think clear that the former case implies, with the assumption of an indefinitely high survival expectancy, a massive per capita monetary gain, while the latter a comparatively extremely low one.

BIBLIOGRAPHY

Aranyosi, István A., 2004: 'Scientistic *Weltanschauung*, Cyborg Dignity, and Cybernethics' (MS).

Bostrom, Nick, 2002: 'Existential Risks. Analyzing Human Extinction Scenarios and Related Hazards', *Journal of Evolution and Technology*, *Vol. 9.*

Bostrom, Nick, 2003: 'Are You Living in a Computer Simulation?', *Philosophical Quarterly, Vol. 53,* No. 211*,* pp. 243-255.

Carter, Brandon, 1983: 'The anthropic principle and its implications for biological evolution' *Phil. Trans. R. Soc*., A 310, pp. 347-363.

Chalmers, David J., 2003: 'The Matrix as Metaphysics', http://www.u.arizona.edu/~chalmers/papers/matrix.html

Epstein, Joshua M. and Axtell, Robert T., 1996: *Growing Artificial Societies. Social Science from the Bottom Up*, Brooking Institute Press, MIT Press.

Leslie, John, 1996: *The End of the World*, London: Routledge.

Levin, Harold L., 1996: *The Earth Through Time,* 5th edition, Saunders College Publishing.

Smith, Quentin, 1998: 'Essay on Leslie's *The End of the World*', *Canadian Journal of Philosophy*, *Vol. 28*, No. 3., pp. 413-434.