APOCALYPSE NOW DOES *THE MATRIX*: ANTHROPIC ADVENTURES FROM DOOMSDAY TO SIMULATION.

Alasdair Richmond

Following on from Nick Bostrom's discussion of the Doomsday argument, Alasdair Richmond considers how anthropic reasoning can lead from Doomsday to some odd conclusions about computation and our place in reality.

The philosophy of explanation can yield strange arguments and this paper looks at two of them, the Doomsday and Simulation Arguments.

One popular view says a good explanation should make whatever you're trying to explain appear more probable, or typical, or more the sort of thing you would expect. Applying this principle, all else being equal, you should expect to be in a fairly probable or typical location for creatures like yourself. For example: carbon chemistry is the largest single branch of the subject. Why? Carbon offers more bonding-opportunities than any other element. If you're going to build complex physical structures, having something like carbon around seems a good start. If observers like us require complex structures (brains, nervous systems, etc.), and carbon is better for building such structures than anything else, it's not surprising we find ourselves living somewhere carbon can exist. Likewise, if observers like us only thrive in a restricted range of temperatures, it's not surprising you and I are (respectively) reading and typing this article on a temperate planet and not (e.g.) inside a star. (If you're reading this article inside a star, please let me know.)

Suppose you think there's a link between (e.g.) being carbon-based and being able to carry out conscious functions. This theory has the advantage that it would make your location as a carbon-based observer more probable, or typical. Likewise, we know the human body needs metals like iron and copper. Seemingly metals are only made in bulk inside stars, and hence only get distributed when their parent stars grow

https://doi.org/10.1017/S1477175600002955 Published online by Cambridge University Press

old and explode as supernovae. So, metal-requiring organisms like us should only expect to live after several generations of stars have passed. Observe the night sky and you'll find stars of all ages. We don't live in the universe's early days when stars were new and of similar ages. And this too isn't coincidental. The early universe was short on metal-scattering supernovae, and hence wasn't friendly to creatures like us.

The above are 'anthropic' arguments. Physicist Brandon Carter coined the name 'Anthropic Principle' to denote the restrictions that our nature as observers set on the kind of physical conditions we are likely to observe. Anthropic effects appear everywhere from cosmology to chemistry. (N.B. 'Anthropic' involves no special reference to humanity – arguments like the above apply to any observers.) Besides physical conditions, there are anthropic reasonings about our location in time, viz. the 'Doomsday Argument'. This argument (also inspired by Carter) says: taking into account your location in human history, you should look with greater favour on the hypothesis that human history is almost over.

How could such a controversial conclusion follow from considering your location in history? It seems likely that the population explosion of the last few centuries has meant that a fair percentage of all the people who have ever lived are alive now. Some estimates suggest our contemporaries may be 10% of the all-time human population. If humanity became extinct tomorrow, our contemporaries will also be about 10% of all the people there *ever will have been*. Imagine every human receives a numbered birth rank, (so the first human born has birth-rank #1, the second birth-rank #2, etc.). If roughly 10% of all people who have ever lived are alive now, and the world's present population is c. six billion, then all living humans have birth-ranks in the vicinity of sixty billion. Imagine humanity survives until there comes a time when the all-time human population is six trillion, (i.e. one hundred times bigger than now). In that case, humans with birth-ranks at (or below) sixty billion will be unusually early – occupying only the first 1% of the all-time human total. However, if we all became extinct overnight, then humans with birth-ranks at (or below)

https://doi.org/10.1017/S1477175600002955 Published online by Cambridge University Press

sixty billion would be all of humanity. On one hypothesis ('Doom Deferred'), all humans up to now are only a fraction of all humanity; on another hypothesis ('Doom Soon'), we are practically all the people there will ever be.

Doomsday's great exponent, John Leslie, illustrates it thus: write your name on a slip of paper and drop it into an urn. Shake the urn and draw out names. You have two theories about how many names the urn contains: ten names or a million. Assuming the draw was random, which theory should you favour if your name is drawn third? Surely the 'ten names' theory – a random draw from only ten names seems much more likely to yield your name quickly than a random draw from a million names. Note the precise numbers involved don't matter – in all cases, you should favour the lowest population consistent with your evidence. (See Leslie's *The End of the World: The Science and Ethics of Human Extinction*, 1998.)

Doomsday makes two key assumptions: i) your birth-rank is randomly selected, and ii) the probability of your receiving the birth-rank you do is inversely proportional to the postulated all-time human population. Thus, if you think the all-time human total will be six trillion, you should think your having your particular birth-rank is a hundred times less likely than it would be if the all-time human total is sixty billion. (Your birth-rank can only equal, or be less than, the birth-rank of the very last human. If you're human, you can't be born *after* the last human ever born.)

Note the importance of randomness – the argument collapses if you think there's something fishy about the draw so your name is more likely than others to appear. Like other anthropic arguments, Doomsday assumes you're pretty unexceptional, considered *qua* random human being. Of course, we're all unique in various ways, but we're also fairly typical in various ways too. The point is: we tend to favour explanations that make our location or history unexceptional rather than exceptional. An example: if mind is purely material, it's conceivable a mind like mine could spring into existence fully-formed through pure happenstance. Maybe lightning struck some chemicals dumped in a swamp and triggered

https://doi.org/10.1017/S1477175600002955 Published online by Cambridge University Press

spontaneous cellular activity that issued in me, complete with false memories of a life I never had. This non-evolutionary 'swamp man' story 'explains' my physical and mental make-up. However, it's a poor explanation that makes me so bizarrely atypical. All else being equal, I won't buy the 'swamp man'story unless I find powerful evidence in its favour. Hence, Doomsday is *not* saying we should treat ourselves as random, undifferentiated humans in every respect. Rather, Doomsday argues anthropically: favour explanations that make you probable unless you've very good reason to think otherwise. Doomsday lecture-audiences often respond 'But I'm unique'. Indeed, experience suggests 68% of all people first exposed to Doomsday retort by asserting their own uniqueness, thereby belying their own uniqueness. (Operational self-refutation?) Being a beautiful snowflake of a person, unique and irreplaceable, doesn't immunise you against anthropic effects. Hands up everybody who's argon-based... No takers? You may be more typical than you'd like to think.

Like all anthropic inferences, Doomsday does not ask you to discount existing evidence so your location may appear more probable. I live in Scotland, whose population the 2006 census put at 5,116,900. China's population is estimated at 1,321,851,888, (July 2007, http://wikitravel.org/en/China). Thus, China's population is (roughly) 258.3 times bigger than Scotland's. Should I conclude I actually live in China? No, because I would then have to discount much of my existing empirical evidence, (e.g. my office's view of Edinburgh Castle). Anthropic reasoning does not say: think yourself probable at all costs – rather, it says: all else being equal, favour hypotheses that make your existing evidence more probable rather than less.

Doomsday is not refuted by saying that earlier historical periods might have made similar inferences. Suppose a centurion in Eboracum (York) c. AD 200, reasons thus: 'World population has doubled since 500 BC, the Antonine Wall's defunct, you can drive chariots through bits of Hadrian's Wall, our technological lead over the Persians is narrowing, the Picts and Teutons won't stay quiet, new catapults and ballis-

https://doi.org/10.1017/S1477175600002955 Published online by Cambridge University Press

tae keep proliferating; overall I give humanity until AD 300 at most.' Our centurion is surely wrong – centuries later, humanity survives. Doesn't this refute Doomsday? Two points: i) Like Hume's Indian Prince, the centurion may reason correctly but to a false conclusion. We know the centurion actually is an unusually early human, and unusual people can easily think themselves more typical than they are. (If you win a lottery, don't assume everybody won.) ii) The centurion's evidential basis is not ours, so we can legitimately give extinction much higher starting-probability. Biochemical weapons and multi-megaton warheads threaten slaughters orders of magnitude beyond anything ancient Rome managed. (Although what befell Carthage was bad enough.)

A Doomsday critic (I name no names) says Doomsday should only trouble atheists, because theists expect eternal life to succeed (Earthly) extinction. However: i) This is no sort of refutation – an injunction not to worry about extinction is not a challenge to Leslie's reasoning; and ii) theists can be just as Doom-phobic as anybody else. Maybe anyone dying in mortal sin is damned. Extinction, especially if rapid, might damn unshriven billions who might otherwise be saved – perhaps including you. Extinction followed by mass damnation is not a consoling prospect.

So, in a nutshell, Doomsday says: once you take into account your location in human history, you ought to believe extinction is more probable than you first thought. Doomsday does not specify *how probable* extinction is – rather than derive a specific probability (e.g. 96%) for extinction, Doomsday says you should increase whatever your starting-probability happens to be. Thus, if you initially think Doom is very unlikely, you might still think that even after considering Doomsday Arguments. So, Doomsday is *not* a prophecy of irrevocable and inevitable doom – you can accept Leslie's reasoning and still think humanity has a long future. Likewise, you might reject Doomsday (e.g. because you reject the probability ideas above) but believe our days are numbered on other grounds, (e.g. ozone depletion, nuclear proliferation, galloping obesity, bird 'flu, Pop Idol – pick your own Armageddon).

https://doi.org/10.1017/S1477175600002955 Published online by Cambridge University Press

But there are odder anthropic horizons yet. Many philosophers think the functional aspects of the mind are constitutive of consciousness. Provided a system instantiates the right functions, it can be fully conscious, regardless of its physical composition. Not what minds are made of, but what they do, is the key. Perhaps mind is to brain as computer programmes are to their supporting hardware, so you can run the same consciousness on different computational vehicles. This 'substrate-independence' view of mind implies you may not be able to deduce much about your substrate from the mere fact of being conscious. However, suppose you don't know what substrate you have but think 99% of observer-substrates are carbon-based. You should accordingly think your substrate is carbon-based, unless you have good evidence otherwise. Likewise, if you believe most observers have sulphur-based substrates and you don't know what kind you have, expect to be sulphur-based. (You could of course be wrong but we're discussing what beliefs you *ought* to have, or are *justified* in having. Alas, false beliefs can be justified.)

Suppose you're a functionalist, who also thinks advanced technology might permit fully-conscious simulations of human consciousness. Let's call fully-conscious simulations 'Sims'. (Cf. Brian Weatherson, 'Are You A Sim?', *Philosophical Quarterly*, 2003.) Computing power has grown for decades. In 1998, I stored my PhD. on a 4-megabyte drive; in 2007, I save these words on a 12-gigabyte drive – my PC-storage has risen 3,000-fold in nine years. We seem nowhere near the theoretical maxima for computing power, speed or efficiency yet. Advanced civilisations may command computing resources vastly beyond ours. (Cf. star-system-sized 'Matrioshka brains' imagined by futurologists.) What might such 'posthumans' do with such computational power? Judging by human experience, they might run many Sims. The computational requirements for running Sims must be huge but then so may be posthuman resources. Assuming our recreational habits are typical of computer-using beings (anthropic reasoning again), posthumans probably run all manner of simulations, with as much depth and complexity as their resources allow.

https://doi.org/10.1017/S1477175600002955 Published online by Cambridge University Press

If your experience is not an infallible guide to your substratum, you might be a Sim and not know it. Substratum-discoveries need more than mere introspection. You can't intuit that you're currently running on a carbon-based architecture – if indeed you are. If you believe a) there are many more Sims than non-Sims and b) your evidence doesn't reveal whether or not you're a Sim, then you ought to believe you're a Sim, i.e. you should favour the hypothesis that Sim-hood is where you're at right now. This is the crux of Nick Bostrom's 'Simulation Argument', ('Are We Living in a Computer Simulation?', *Philosophical Quarterly*, 2003). Bostrom maintains his argument differs in a key respect from Doomsday: if we don't know whether or not we're Sims we can regard ourselves plausibly as random consciousnesses, but we do know our birth-ranks and hence cannot view ourselves as random humans.

Note Bostrom is not saying we're probably Sims. Rather, he says: *if* you believe in functionalism *and* believe Sims outnumber non-Sims, *then* you should believe you're probably a Sim. However, Bostrom's reasoning is compatible with other possibilities: 1) posthumans are rare (e.g. extinction prevents most species reaching posthumanity); 2) posthumans run few Sims; 3) functionalism is false. But recall the argument's anthropic inspiration: if we attain Sim-technology, this would tell heavily against options 1-3. Assuming we're typical computer-users, our running Sims would be powerful evidence that we are ourselves Sims. (This outcome would support functionalism, while telling against both the rarity of posthumans and posthuman reluctance to run Sims.) So, if we run Sims, we might suspect that reality is multi-layered, with our Sims being Sims run by Sims (i.e. us), whose simulators in turn may be Sims. (If functionalism is right, a correct simulation of a Sim is itself a Sim.) Presumably, if the master-simulators at the bottom of the pile have only finite resources, this regression couldn't be infinite but it could be deep nonetheless.

Just as you can accept Bostrom's argument but reject your Sim-hood, you can reject Bostrom's argument but still believe you're a Sim. Maybe you've observed a computational 'glitch' in your environment or you think your simulators have contacted

https://doi.org/10.1017/S1477175600002955 Published online by Cambridge University Press

you directly, (although don't embrace either explanation until you've eliminated some alternatives first). Bostrom's conclusion is actually *disjunctive*, (i.e. an either/or choice between various options). An either/or conclusion is something Simulation has in common with Doomsday. As Leslie insists, despite its name, Doomsday is compatible with outcomes besides extinction. Maybe humanity survives but in so different a form that our descendants don't belong in the same reference-class as us. It's not clear that the all-time census of people should include our hominid ancestors – perhaps our near-descendants make some cognitive breakthrough which puts them into a different bracket again. Our descendants might upload their minds onto computers and thereafter acquire new cognitive abilities simply by bolting-on new modules. (For runaway computer-intelligence, see I. J. Good, 'Speculations Concerning the First Ultraintelligent Machine,' *Advances in Computers*, 1965.) Combining Doomsday with functionalism might portend, not extinction, but extended human/machine symbioses. Some 'Transhumanists' cheerfully anticipate becoming effectively immortal via such fusion. However, Transhumanist practices may themselves be dangerous. If we lavish too many resources on indefinitely prolonging a few at the expense of securing basic needs for the many, extinction may claim us all. Let's worry about physical immortality once everyone has drinking-water. Why should my bodily life extend forever? Remember the anthropic moral: we're probably more typical than we think. Being human, carbon-based and mortal offers challenges enough. (Any ontological condition good enough for Martin Luther King and Uma Thurman, to name but two, seems good enough for me.)

Alternatively, joining Doomsday with Simulation might suggest human extinction has already occurred, i.e. we aren't carbon-based after all (and maybe never were). Computer intelligence may have evolutionary advantages – more tolerant of radiation and extremes of temperature, less reliant on air and water. If you think computer intelligences predominate, and/or most species only spend part of their history embodied, you might be sceptical about being human. Or maybe such

https://doi.org/10.1017/S1477175600002955 Published online by Cambridge University Press

'Apocalypse Past' speculations are a *reductio ad absurdum* of functionalism.

Generalising from human computer-use prompts humbling thoughts. The Search for Extraterrestrial Intelligence (www.seti.org) offer screensavers that let your computer devote spare computing capacity to sifting astronomical data. Our lives may be some posthuman screensaver-equivalent, run while our creator grabs a post-coffee. Ignominious Doom looms if our creator's break finishes and all this spare computing capacity goes back to work. Combining Doomsday and Simulation, maybe most Sims live in the last splurge of Sim-activity just before it all switches off... (Compare desperately e-mailing just before your Internet access runs out or frantically finishing off an exam against the clock.) If a conscious being measures time-flow by the number of processes it carries out per temporal unit, there seems no reason why Sims couldn't experience subjective lifetimes in mere seconds of simulator-time. We might live 'longer' lives than our creators even if there's no point in time at which we are alive and they aren't. (In the limit, Sim-time might even be infinite but simulator-time finite, making us immortal creations of mortal 'gods'.)

Philosophy needs thought-experiments, and science fiction is a rich source. However, while I love SF and shamelessly used the *Matrix* to grab your attention, I find the *Matrix* trilogy dull, superficial and unpleasant – and significantly less interesting than the Sim-world postulated by Bostrom. (*Apocalypse Now* though is magnificent – not very anthropic but an irresistible name-drop.) In the *Matrix* trilogy, virtual reality simplifies things rather than the reverse: reality may have two levels but it's pretty easy to tell a) which level is which and b) who the bad guys are. I also dislike the trilogy's gestures at mysticism. Proper mystics are well-rounded people, not gun-happy narcissists. Despite the trilogy's 'spiritual quest' noises, all it takes to solve Matrix-problems is squaring your jaw and hitting things. Reality (virtual or not) just ain't like that. If you prefer thought-provoking virtual reality fictions, I recommend Christopher Priest's novels *A Dream of Wessex* (1977) and *The Extremes* (1998), and David Cronenberg's film *eXistenZ*

https://doi.org/10.1017/S1477175600002955 Published online by Cambridge University Press

(1999). (Priest especially makes genuinely unsettling and inventive use of simulated worlds.)

Are Sim-worlds like the Matrix? Yes and no. The trilogy's 'heroes' are not really Sims *per se* but normally-embodied persons who are fed a systematic simulated world. The purely computational Smith is more like a true Sim. (One rare interesting twist in the trilogy's later instalments is Smith uploading himself onto human brains, making him a carbon-based Sim.) The *Matrix* portrays one restricted kind of Sim-world. Let's call *unconscious* simulations 'simulacra', so simulacra are computer-generated zombies, *not* Sims. In the *Matrix*, a few 'real' people (and machine-possessed baddies like Agent Smith) live amid a large number of simulacra. Many Matrix simulated people are seemingly window-dressing. There's no reason to assume most Sims will live in Matrix-style simulations with only few Sims to many simulacra. Bostrom's argument suggests an altogether richer, multi-layered reality whose different levels might be genuinely diverse yet filled with conscious beings throughout. For functionalists, a proper simulation of consciousness is itself conscious, and having a computational substratum is no reason why you can't be fully conscious. Thus, we should respect Sims as persons, even if Sims make backup copies of themselves. (Sympathy for Agent Smith? He's just trying to survive, albeit aggressively.) This last isn't as frivolous a point as it may initially sound. Functionalism means backup copies might be conscious too. To deny personhood and its concomitant rights to a conscious being solely because it has a substratum unlike yours is effectively tantamount to racism or speciesism.

Matrix fantasies seem psychologically and philosophically unhealthy. ('I may be obscure in this delusory world but in reality-at-large, I'm a shade-wearing superbeing – soon, these simulacra shall call me Messiah.') Solipsistic power-fantasies often take their owners off life's stage but sadly, such owners don't always depart alone. Some *Matrix* fans once told me they believed 99% of (so-called) 'other people' were simulacra. Not knowing what response to make, I didn't reply, but pragmatism suggests: On meeting things that behave like people, please

https://doi.org/10.1017/S1477175600002955 Published online by Cambridge University Press

assume they *really are people*, with inner lives and agendas as valuable as yours. Treating simulacra like persons may be foolish, but treating persons like simulacra is criminal. If in doubt, treat the person-like as persons. Whatever blunts our sensitivity to the human seems risky. (An argument against virtual reality – not to mention android stress-dolls and sex-toys?) Here multi-layer Bostrom-worlds differ significantly from two-level Matrices. In Bostrom's world, no level can be sure it's the basement so all levels face the possibility of censure from simulators further down. Thus, in Bostrom-worlds, individuals and species alike have incentives to behave morally. So, for several philosophical and moral reasons, I'd rather this was a Bostrom-world than a Matrix.

Finally, do I accept Doomsday or Simulation? No – while I don't accept all the criticisms levelled at them, ultimately I can't buy either argument. (See my 'Recent Work: The Doomsday Argument', *Philosophical Books*, 2006, and 'The Simulation Argument & Simulation Hypotheses', *Philosophy Through Science Fiction*, ed. Ryan Nichols, forthcoming 2008.) The reference-class of human observers must be left deliberately vague in order to support Doomsday conclusions and to preserve our 'randomness' as humans. Most probabilistic inferences seek to make their reference-classes more specific rather than less and to use as much relevant information as possible in delimiting their reference-classes. For some purposes, we can afford to keep our reference-classes fairly vague; in life-and-death matters, we tend to fix our reference-classes as accurately as we can. The chances of my being carbon-based don't decline if I narrow my reference-class to Scottish life-forms. However, my being Scots may affect my risk of coronary heart-disease – in this case, a more inclusive reference class might sharply change the probabilities, and hence my indicated survival-strategy. Doomsday's probability assumptions also seem suspect, especially that one's birth-rank gets less probable as the overall postulated population rises. This strategy is one way of awarding birth-ranks probabilities but other strategies seem equally plausible *prima facie*. Maybe a larger human race affords more possibilities for being human,

https://doi.org/10.1017/S1477175600002955 Published online by Cambridge University Press

and thus raises your birth-rank's probability. (Although neither probability-setting strategy seems compelling taken singly.)

Overall, Doomsday seems both more anthropic and more robust than Simulation. (Even discounting my strictly armchair assessments that human extinction is sadly plausible but 'Matrioshka brain' techno-optimism is so much fiddling while Rome burns.) Anthropic reasoning tries to counteract the seemingly innate human tendencies to anthropocentrism and self-inflation in assessing our place in nature, whereas Simulation's picture of a stage-managed world is counter-anthropic and almost calculated to put us back in the centre of existence. The leap from anthropic considerations to a multi-layer hierarchy of simulating Sims seems suspicious. Ramifying reality thus seems deeply inflationary if undertaken on so slight an evidential basis, and drastically restricts the appeal of one leg of Bostrom's disjunction. (It might be a different story if we started getting messages from our simulators or somehow detected a computational substratum beneath our world's sensory appearances. However, I suspect nothing of the kind has been observed hitherto and that it would take astonishingly good evidence to make such hypotheses compelling. A neo-Hume would have a field day with testimony to such events.)

Doomsday uses a reference-class that seems vague but broadly justifiable – after all we already have to use a rough-and-ready concept of 'human being' for many purposes. However, the Simulation reference-class seems both vague and contrived – as though the argument pre-dated its reference-class and not *vice-versa*. Simulation also invites questions about how belief supervenes on evidence and how Sims might keep scepticism at bay. The smooth running of a Sim-world seems to leaves Sims in a position compounded of large measures of contrivance, surveillance and epistemic luck. Anthropic reasoning is subtle and often highly persuasive but it has yielded many applications more compelling than Doomsday and Simulation.

Alasdair Richmond is Lecturer in Philosophy at Edinburgh University.

https://doi.org/10.1017/S1477175600002955 Published online by Cambridge University Press