Barry Dainton

On Singularities and Simulations

If we arrive at a stage where artificial intelligences (or AIs) that we have created can design AIs that are more powerful than themselves, and each new generation of AI rapidly creates still more powerful AIs, then the 'intelligence explosion' — or *singularity* — foreseen by Good, Vinge and others could easily become a reality. Since the arrival of superintelligent machines would be a momentous, world-changing occurrence, we would be wise to consider how best to deal with this eventuality should it occur; we should also attempt to ascertain whether the singularity is as imminent as some of its proponents maintain. David Chalmers' 'The Singularity: A Philosophical Analysis' contains much that is valuable on both fronts. With regard to the key issue of whether a singularity is possible at all, I think Chalmers is right in saying that it is certainly not out of the question. As for how to minimize the dangers posed by an emergent superintelligence, the measures Chalmers proposes — implanting the right values, isolating the first super-intelligent systems in virtual universes — look to be promising avenues.

My focus in what follows will be on some of the consequences of a computer-based intelligence explosion, assuming we can survive it. The combination of superintelligence and massive power will make it possible for computers to create and sustain virtual environments of a size and complexity that is way beyond anything we are currently capable of devising. Will it be possible — or desirable — to 'upload' ourselves into these virtual worlds? Chalmers has interesting things to say on this issue; I will be suggesting a slightly different take on it.

Correspondence:
Email: bdainton@liverpool.ac.uk

Journal of Consciousness Studies, **19**, No. 1–2, 2012, pp. 42–85

The existence of these virtual worlds leads to another question: might it not be possible — or even probable — that we ourselves are among their (virtual) inhabitants? This issue has been discussed fairly widely in recent years. I will be proposing some relevant (and hopefully, useful) distinctions and offering some reflections on it. But first, by way of a preliminary, a brief excursus into science fiction proper.

1. New Visions of Heaven

In his 1997 novel *Diaspora*, the Australian science fiction author (and computer programmer) Greg Egan describes a future in which humankind has split into three groupings of very different kinds of being, two of which are decidedly 'post-human':

> *fleshers* These are humans of the flesh-and-blood variety. Some fleshers are biologically very similar to ordinary 21st century people, others have modified their genes so as to have more-than-ordinary-human attributes (e.g. greater resistance to disease, longer life-spans, higher intelligence, the ability to breath underwater, etc.); some — the 'dream apes' — have deliberately reduced their cognitive capabilities in order to commune with nature in a deeper way.
>
> *citizens* Software-based subjects, entirely lacking in flesh or blood but fully conscious, who live in virtual environments or *polises* sustained by powerful computers. There are a large number of polises, each with its own charter setting out the distinctive approach to living and the external world to which its citizens (and the polis's AIs) are committed. In effect, each polis is a distinct civilization, containing thousands (or millions) of inhabitants. Some polises largely devote themselves to maths and science, others to art; some are largely solipsistic, ignoring the external world, others are more outgoing and engaged; citizens can move between polises as their outlooks and orientations change. Some citizens started off their lives as embodied humans and entered their polis by uploading, others are polis-born.
>
> *gleisner robots* These are also post-humans whose minds are running on computers, but these computers are housed in non-biological (robot-like) physical bodies. Gleisners are more resilient than flesh-and-blood humans, but can interact with the physical world in the same sorts of ways.

The main action of the novel is set a fair way into the future, towards the end of the 29th century, and by this time the population of the polises vastly outnumbers that of the gleisners and fleshers, but humanity began its transfer into the realm of the virtual (the *Introdus*, as it's called in the book) a good deal earlier, in the mid-late 21st century — or so we are told.

This is all nothing but a fiction, of course, but if the technological capabilities were to arise, the idea that humankind would divide (or 'speciate') along roughly these lines strikes me as very plausible. Since there are already some people who are eager to have themselves uploaded — anyone sceptical of this should spend a few minutes at any of several post-humanist websites — there would obviously be *some* people who would be more than happy to go through with it soon as the technology becomes available, and more would no doubt follow when the technology matures and becomes more reliable. And if life in a virtual world should turn out to resemble that described in *Diaspora* the trickle of uploaders might well become a flood, for Egan's citizens enjoy a lifestyle which, in many respects, is highly enviable. Those citizens who want to devote themselves to pure research (in maths, physics, philosophy, whatever) can do so, albeit with their intellectual abilities and powers of concentration considerably augmented. But those who want to devote their lives entirely to the creative arts are free to do so, and ditto for those who want to spend their time partying or socializing, or those who prefer a balance between intellectual and other pursuits. Citizens who are curious about life as an embodied human need not worry: they can spend as much time as they like in fully accurate simulations of pre-upload life; those who are curious about what life would be like with a different kind of *mind* can have their own personalities altered, temporarily or permanently. This all sounds highly appealing, and I haven't yet mentioned the freedom from ailments and ageing, and the promise of immortality (or close to it) for those who want it.

However, while the prospect of all this would no doubt appeal to a great many, there would inevitably be those who would be reluctant to take their leave of ordinary material reality, and for a variety of reasons. For some, the idea of trading a real living body for (what is arguably) no more than the illusion of one would simply be too distasteful to contemplate. Others might instinctively recoil from the thought of trading a real physical environment for a virtual one; perhaps some would be fearful that virtual life would be in some way less vivid — *less real* — than an ordinary life.[1] Some might be swayed by religious doctrines, while others might be swayed by more philosophical

[1] Such sentiments are certainly not unknown amongst fleshers in *Diaspora*. With life on Earth threatened by an imminent gamma ray burst, the fleshers are offered the opportunity of uploading to escape the danger; one responds thus: 'You are shameless. We expect no honour from the simulacra of departed cowards, but will you never give up trying to wipe the last trace of vitality from the face of the Earth? … Did you imagine that a few cheap, shocking words would send us fleeing from the real world of pain and ecstasy into your

considerations. Can we really be sure that the inhabitants of these computer-sustained virtual worlds are *conscious*? Even if we have good grounds for supposing they are, is the uploading process genuinely person- or self-preserving? It is far from obvious that it would be. Here is what happens to a character in *Diaspora* (Orlando) during the initial (highly destructive) phase of the upload process:

> Waves of nanoware were sweeping through Orlando's body, shutting down nerves and sealing off blood vessels to minimize the shock of invasion, leaving a moist pink residue on the rubble as flesh was read and then cannibalized for energy. Within seconds, all the waves converged to form a grey mask over his face, which bored down to the skull and ate through it. The shrinking nanoware spat fluid and steam, reading and encoding crucial synaptic properties, compressing the brain into an ever-tighter description of itself, discarding redundancies as waste.
>
> Inoshiro stooped down and picked up the end product: a crystalline sphere, a molecular memory containing a snapshot of everything Orlando had been. (Egan, 1997, p. 110)

The envisaged storage-crystal may contain an accurate recording of Orlando's psychology, one which at some future time can be brought to (virtual life). But is it really possible for Orlando to survive having his brain taken apart — *boiled*, by the sound of it? Will the virtual-Orlando be *Orlando*, or merely a facsimile of him?

2. Issues of Experience and Identity

Chalmers discusses these and related issues in the latter stages of 'The Singularity'. If current research into AI does lead to the emergence of a superintelligence (of the A+ or A++ varieties), then assuming we survive we will have to decide how to respond.[2] Although some might prefer to interact as little as possible with the superintelligent systems, this isolationist (flesher-like) stance would not be the preference of everyone: many would undoubtedly seek to interact with them, albeit with differing degrees of intimacy. But there is a problem: it is very likely that A++ systems will function many times more quickly than ordinary humans — in the fictional the *Diaspora* universe, for example, the subjective time of the citizens runs some eight hundred times

---

nightmare of perfectibility? … Why can't you stay inside your citadels of infinite blandness, and leave us in peace.' (Egan, 1997, p. 92).

[2] In Chalmers' terminology, an 'AI' is an artificial intelligence of roughly human level, an 'AI+' system is an AI that is *more* intelligent than the most intelligent human, whereas an 'AI++' system is an AI that is at least as far beyond the most intelligent human as the latter is beyond the typical mouse; A++ systems possess *superintelligence* (2010, p. 11).

faster than that of the fleshers, and interaction between the two camps is infrequent for precisely this reason. So those seeking to interact frequently and efficiently with a superintelligence will need to boost the speed of their mental processes. While brain enhancement technologies may be of some assistance in this regard, they are unlikely to take us very far, and I think Chalmers is right when he suggests that in the long run 'if we are to match the speed and capacity of non-biological systems, we will probably have to dispense with our biological core entirely' (2010, p. 20). So to integrate with the AI+ systems we will have to transfer our own minds into non-biological computers in the manner of Egan's citizens: we will have to upload our minds.

Assuming the medical and technological obstacles could be overcome — superintelligent computers could surely help with these — anyone contemplating undergoing an upload is confronted with the two philosophical questions we have just encountered. First of all, will the product of an upload be *conscious*? More generally, are the sorts of computers which are capable of running AI++ systems capable of sustaining communities of beings who are conscious? Second, is uploading a process one could survive? If I were to undergo an upload process, would the resulting subject be *me*, or another person entirely, one who happens to closely resemble me psychologically?

With regard to the issue of whether an upload (or an A++ system) could be consciousness at all, Chalmers admits that the issue is complicated by the fact that our understanding of consciousness is so poor, but he also thinks that a strong (if less than conclusive) case can be made for holding that consciousness is an organizational invariant, i.e. that systems possessing the same patterns of causal organization will instantiate the same types of conscious states, irrespective of whether the organization is implemented in neurons, silicon, plastic, or any other substrate. And because of this he is reasonably confident that suitably programmed computers can be conscious — and hence generate virtual worlds whose inhabitants are as conscious as we are. Although I am less confident than Chalmers with regard to the strength of the case for holding consciousness to be an organizational invariant, I am in full agreement with him that there's a great deal that we don't yet know about the physical underpinnings of consciousness. As a consequence, I do not think we can sure computers could *not* be conscious. For present purposes I will work on the assumption that they can be. What I want to focus on is the personal identity issue. Which sorts of upload-process are identity-preserving? Can we be confident that *any* form of uploading is truly person-preserving?

Chalmers' nuanced and interesting discussion of this issue in *The Singularity* is (to my mind) largely very plausible. But it is also very cautious: in Chalmers' eyes, much remains unsettled. On what he labels the 'optimistic view', uploading can be survived — at least if done in optimal ways — but on the 'pessimistic view' the process is fatal. Uploading itself may come in many very different forms, and to make matters concrete Chalmers orientates much his treatment around two specific forms which he labels *destructive* and *gradual*.[3] The process Orlando underwent in the passage cited above is an instance of destructive uploading. In such cases all the information relevant to creating an accurate psychological copy of an ordinary human subject is extracted from their brain by a sophisticated scanning process and safely stored, but the subject's brain is destroyed as a consequence; at some later time the stored information is used to create a psychologically similar subject in a virtual world. Since for a period (even if only a short one) the subject ceases to exist, then even if we assume they re-enter existence when the replica is created, a destructive upload does not preserve full mental continuity: at the very least the subject will lose consciousness for a while. As construed by Chalmers, *gradual* uploads do not involve any such rupture in the mental lives of those who undergo them; most notably, if a gradual upload is carried out quite quickly — over a matter of minutes or hours — they need not disrupt the continuity of the subject's consciousness: it is perfectly possible for the subject to remain fully awake and aware throughout the procedure. But although gradual uploads are non-disruptive in *this* way, in other respects they can be highly damaging. The nano-replacement procedure Chalmers describes — which involves the gradual replacement of all a subject's neurons by silicon-based devices — totally destroys the biological brain of the subject who undergoes it.

Since Chalmers believes that the uninterrupted continuity of consciousness is a particularly reliable guide to personal identity (2010, pp. 54, 60) he holds that anyone who undertakes a gradual upload should be confident that they will survive the process. Chalmers also tells us that while he is more sympathetic to the psychological account of personal identity (which permits successful destructive uploading)[4]

[3] In the interests of brevity I will not be discussing Chalmers' rather more speculative *reconstructive* upload process.

[4] In Parfit's widely-used terminology, it is *wide* psychological continuity which renders processes such as destructive uploads and teletransportation survivable; on this view, persons P_1 at t_1 and P_2 at the later time t_2 are one and the same person if P_2's psychological states (beliefs, memories, intentions, desires, etc.) are both similar to those of P_1, and

than the biological account (which doesn't), he isn't confident that the psychological account is correct, and hence 'I am genuinely unsure whether to take an optimistic or pessimistic view of destructive uploading. I am most inclined to be optimistic, but I am certainly unsure enough that I would hesitate before undergoing destructive uploading' (2010, p. 50).

Why so much uncertainty? Although a destructive upload preserves psychological continuity (at least of the wide variety), Chalmers doesn't think it is intuitively clear that the process is in fact person-preserving, and this casts some doubt on the psychological account itself. In addition, there are two alternative views of personal identity which deliver different verdicts on destructive uploads, and while Chalmers isn't sure that either of these alternatives is correct, he is not sure they are *not* correct either. The first alternative can be summarized thus:

> *The further fact view*: knowing all the facts about personal physical and psychological facts in a given case doesn't provide one with knowledge of the facts about personal identity. If, for example, there are primitive immaterial substances (or souls), and we are identical with these substances, then the facts about personal persistence would be determined by the facts about these, rather than any facts concerning biological or psychological relationships.

This epistemic gap could have other sources, e.g. our *concept* of personal identity might simply be such that even after all facts about mental and physical continuities are specified, it remains open whether these facts suffice to secure personal persistence in the context under consideration. In any event, Chalmers tells us that he thinks a further fact view *could* be true; he does not think it is ruled out by anything that we know. If a further fact view *is* true, then it is unclear whether destructive uploading is person-preserving or not.

Chalmers calls the second alternative the 'deflationary view'; it is harder to pin down in a succinct formulation, but its main ingredients are as follows:

> *The deflationary view*: we are inclined to think personal identity is more solid and determinate across a wider range of circumstances, both actual and possible, than it really is. In puzzling cases — such as destructive upload, or teletransportation — where it is not intuitively clear whether the original person survives or perishes even when all the information relating to mental and physical continuities is known, there

causally dependent on those of P_1 in a suitably direct way — a way which does *not* require sameness of brain, or indeed, a continuously existing mind of any kind. It is this last proviso which allows people to survive being reduced to passive collections of data.

> simply *is* no fact of the matter, whether of the ordinary or 'further' variety, as to whether the process in question is person-preserving or not. In the absence of identity-facts, when deciding what to make of such cases we have no option but to focus on the facts relating to the mental and physical continuities which do exist, and form a view as to their importance: how much of what matters in a life do they preserve?

Intriguingly, Chalmers suggests that the deflationary approach can be extended to ordinary cases of survival: 'we are inclined to believe in Edenic survival: the sort of primitive survival of a self which one might suppose we had in the Garden of Eden. Now, after the fall from Eden … there is no Edenic survival, but we are still inclined to think as if there is' (*ibid.*, p. 60). In this guise the deflationary view can make uploading seem less unpalatable. Uploading may not ensure perfect Edenic survival, but neither (it turns out) does ordinary life, and so in this sense uploading is not *that* much worse than waking up after a period of dreamless sleep. It is true that destructive uploading does not preserve biological continuity, but it does preserve causal continuity and psychological similarities, which relative to our actual scheme of values carry a good deal of what matters in ordinary survival. And gradual uploading, which does not disrupt the continuity of consciousness, preserves a very great deal of what matters.

3. Uploading: Another Perspective

The relationship between between the continuity of consciousness and personal identity is of a distinctively intimate sort. Letting your imagination roam far and wide, can you envisage a state of affairs, *any at all*, in which your current stream of consciousness goes one way, and you go another? I suspect not. Provided our consciousness flows smoothly on — i.e., provided the experienced succession of bodily feelings, perceptual experiences, thoughts and mental images that is characteristic of our ordinary streams of consciousness is uninterrupted — we can be certain (or as certain as we can be of anything) that we ourselves are continuing to exist, irrespective of what else is happening to us.[5] It matters not if the neurons in our brains are replaced with silicon surrogates; provided we continue to experience, we continue to exist. Likewise for psychological manipulations (or advanced brainwashing) which alter our memories or beliefs: these

[5] If you think you have succeeded in imagining a procedure which involves yourself and your stream of consciousness going their separate ways, consider: if you also envisage yourself as remaining conscious throughout the process, then all you have done is imagine your original stream of consciousness smoothly dividing into two, and this isn't the same thing at all.

too we can envision ourselves surviving, provided they do not impact on the flow of our experience.[6]

Chalmers repeatedly says that he thinks that continuity of consciousness is the most reliable guide that we have to the continued existence of a self or person. Yet he is also reluctant to rule out the further fact and deflationary views. This may seem puzzling. For if the continuity of consciousness is sufficient for one's survival, then how can there also still be room for any kind of 'further fact' to play a role? And if continuous consciousness can secure one's persistence in a perfectly secure manner, aren't the claims of the deflationist also undermined? What does Edenic survival have to offer that ordinary survival lacks?

As far as I can see there are two reasons why Chalmers adopts the stance that he does. First, in much of his discussion he works within the confines of the orthodox (essentially Parfittian) view that facts about personal identity are determined by biological and psychological-cum-causal facts. If we take these as our base-level facts, then evidently any facts about the continuity of consciousness will be counted as *further* facts, for the orthodox framework makes no mention of experiential continuities. Indeed, Chalmers acknowledges that if we include facts about the experiential continuities all ambiguities are removed: ''I think it is plausible that once one specifies that there is a continuous stream of consciousness over time, there is no longer really an open question about whether one survives' (2010, p. 60). But this takes us on to the second point. It may be entirely clear-cut that you continue to exist for as long as you are enjoying an uninterrupted stream of consciousness, but what happens when you are no longer doing so? What happens when you lapse into the sort of dreamless sleep that most of us enjoy every night? Unless we opt to say that it is impossible for a person to survive interruptions in their experience (a decision which would drastically shorten all of our lives) then we need a plausible account of the conditions under which distinct streams of consciousness — streams that are separated by periods of time during which their owners are not enjoying any form of experience — belong to the same person or self. Must such streams be generated by the same brain? Must they be associated with causally related psychological systems? Must they be instantiated in the same primitive immaterial substance? Is there some more mysterious ingredient — some further fact — which performs the job? Can we even be sure that our identities *are* preserved through periods of unconsciousness in as

[6] See Dainton and Bayne (2005) and Dainton (2008, chapter 1) for more on this theme.

secure a manner as they are preserved through uninterrupted periods of consciousness?[7] In the absence of a plausible account of how (or even whether) we survive periods of unconsciousness, Chalmers is right to hold that these are open questions.

The difficulty here is a real one, but it is by no means insuperable. On a number of occasions (see Dainton, 2004; 2008) I have argued that there is a natural way to solve the problem of experiential gaps for anyone who wants to take the continuity of consciousness as their primary guide to personal identity. This is not the place for a full rehearsal of this account — I call it the 'C-theory' — but for present purposes a broad overview of it will suffice.

Although some have held that we are identical with our experiences, this is not a very plausible or appealing view; it is more natural to hold that we are *things that have experiences* — or in the usual terminology, we are *conscious subjects*. Whatever else they may be, subjects of this sort typically have capacities to have a range of different kinds of experiences — bodily sensations, mental images, perceptual experience, conscious thoughts, and so forth — capacities which are sometimes exercised (during our waking hours) and sometimes not (when we are unconscious). In our own case these capacities are grounded in our brains, but since for all we know it may be possible for things quite different from a human brain to possess capacities for experience we need a more general term for things thus equipped; let's call them *C-systems*. Under what circumstances do C-systems at different times belong to the same subject? Well, we can say at least this: C-systems which have the ability to produce continuous streams of consciousness belong to the same subject. Since C-systems which are dormant (or unconscious) can still have the *ability* to produce such streams, this criterion applies equally to C-systems which are active and producing experiences, and those which happen to be inactive.

For anyone who takes the continuity of consciousness to be our best and most reliable guide to personal persistence, the notion that C-systems should be assigned to the same subject on the basis of their ability to contribute to single uninterrupted streams of consciousness is the obvious way to go — indeed, what criterion could be more secure or more readily intelligible? It is not difficult to construct a general account of personal or self-identity on this basis. Let us say that two brief C-systems (or phases of such) at neighbouring times are *directly*

[7] Chalmers says he does not endorse, but nor is he entirely unsympathetic with the view that 'we Edenically survive during a single stream of consciousness but not when consciousness ceases. On this view, we may Edenically survive from moment to moment, but perhaps not from day to day' (2010, p. 61).

stream-related if they are either (i) both active, and the experiences they produce form parts of a single continuous stream of consciousness, or (ii) they are not both active, but if they were the experiences they produce would be parts of a single continuous stream of consciousness. Let us call C-systems that are not directly stream-related, but which are linked by a chain of C-systems that are so related *indirectly stream-related.* By way of final terminological move, let us say that a series of C-systems existing at different times are *C-related* only if they are either directly or indirectly stream-related. The C-theory can now be stated in a simple way: C-systems (or phases of such) belong to the same subject (or self, or person) if and only if they are C-related. With this much in place we can opt to identify selves (or subjects) with C-systems — the option I prefer — or merely trace the identity of subjects via their C-systems. For present purposes nothing hangs on which of these options we adopt.[8]

The C-theory provides a simple but effective solution to the problem posed by interruptions in the continuity of consciousness, and it does so by appealing to nothing more than the continuity of consciousness — or at least the potential for it. So far as the intuitive appeal of the account is concerned, the change of focus from actual continuity of consciousness to potentialities for it is of little or no consequence. I have a special concern about the kind of experiences which will feature in my current stream of consciousness as it continues to flow on until I next lose consciousness. Why? Because these experiences will all clearly and unambiguously belong to me, and like most of us I have a primitive and profound concern about the character of my own experiences, particularly those which lie in my future. As is easily seen, a simple but powerful mechanism extends the reach of this self-oriented concern to C-systems and their constituent experiential capacities.

Given that my current stream of consciousness belongs to me — utterly without ambiguity — so too do the experiential capacities which are producing the experiences this stream contains. However, I also possess experiential capacities which are not active and producing experience, but which could be; e.g., if I were to turn my radio on, I would have auditory experiences that I wouldn't have otherwise.

[8] The C-theory as expound at greater length in *The Phenomenal Self* (2008) is a more detailed elaboration of this approach; I also extend the approach to the synchronic case, by construing C-systems-at-time as collections of experiential capacities which are rendered co-subjective by virtue of the ability to produce synchronically unified states of consciousness; I also broach the tricky topic of what to say about branching streams of consciousness (or fission cases), a topic I leave to one side here.

The C-theory easily accommodates experiential capacities falling into this category. By way of illustration, let N_1, N_2, N_3 … be intervals of time of one second duration. (This is just to keep things as simple as possible — the same considerations apply over shorter intervals.) Let us further suppose that my current experiences are taking place in N_1. What should we say about the ownership of the dormant experiential capacities in N_2? According to the C-theory, experiential capacities which exist at adjoining intervals belong to the same subject if and only if they are either (i) active and contributing to a single continuous stream of consciousness, or (ii) they are not all active, but they are such that *if they were*, they would be contributing to a single continuous stream of consciousness. Inactive experiential capacities obviously cannot satisfy condition (i), but they can satisfy condition (ii). Dormant experiential capacities located in interval N_2 belong to me only if they are such that if they were active and producing experiences, these experiences would feature in the direct continuation of my current stream of consciousness (i.e. the phase located in interval N_1).

Should my special self-interested concern extend to these inactive experiential capacities? The answer is clearly in the affirmative. The N_2-phase of my C-system (we can safely suppose) includes a capacity for intense sensations of pain. Considering matters from my present vantage point in N_1, I would very much prefer that this capacity remains inactive; after all, if it is triggered, then the resulting pain-sensation will occur in the very next phase of the stream of consciousness I am currently enjoying. If my present conscious state will flow directly into a state which includes this sensation, without any interruption in experiential continuity, how can I doubt that it will be *me* who feels (and suffers) the pain? This special concern extends to all the other powers which belong to the N_2-phase of my C-system, since these can all influence the character of the direct continuation of *this* stream of consciousness in the immediate future. Hence if during N_1 I turn my radio on, auditory experiential capacities will be activated during N_2, and I will hear sounds I would not have heard otherwise.

Let us next suppose that in the interval N_3 *none* of the experiential capacities which the C-theory assigns to me will be active — that during this period I will be enjoying a few moments of complete unconsciousness. This does not affect the situation in the slightest. As we have just seen, not only do the dormant powers in the N_2-phase of my C-system all unambiguously belong to me, but my primitive prudential concerns inevitably extend to them without diminution or dilution.

The N_3-phase of my C-system may be dormant, but it is nonetheless C-related to the N_2-phase, i.e., if the experiential capacities in these successive phases *were* active, the experiences they produce would form parts of a continuous stream of consciousness. As a consequence, the ownership of the N_3-phase capacities could not be clearer: they belong to me, and they do so in the same utterly unambiguous way as their N_2-counterparts. Given this, my special prudential concern naturally and inevitably extends to them as well. And the same applies to the N_4-phase of my C-system; even if experiential capacities here are all inactive, they have the ability to join with my N_3-phase capacities in the generation of a continuous stretch of experience. This mechanism for guaranteeing sameness of owner (or subject) of successive C-system-phases, and thence transmission of prudential concern, can operate over indefinitely long periods of time — over entire lifetimes.

The C-theory provides us with a clear and informative answer to the question 'What makes it possible for streams of experience separated by a gap in consciousness belong to the same self?', and does so wholly in terms of capacities for experience: earlier and later streams have the same subject if they are C-related. With C-relatedness on hand there is no need to appeal to bodily or material continuity, psychological-cum-causal connections, primitive immaterial substances or some mysterious further fact to explain how experiences separated by gaps in experiential continuity can and do have the same subject.

The C-theory also goes a long way towards undermining the version of the deflationary view according to which our persistence, even in ordinary circumstances, falls short of what it might be — or what we hope it might be. It is not obvious how survival could be better than what the C-theory has to offer, even in Eden.[9]

[9] Following Chalmers' lead I have simply assumed here that the our typical streams of consciousness *do* exhibit a distinctive and robust form of continuity. This is not the place for defense of this *prima facie* plausible claim against those who would deny it — this is a task I have undertaken elsewhere: see Dainton (2008, chapter 3; and forthcoming); see Dainton (2010) for an overview of the debate on this issue. In a different vein, it might be argued that, other things being equal, it is better to spend as much of one's time awake as possible, and so a life which includes periods of dreamless sleep is less good than a life of the same duration which contains no such periods. But even if it is better to be experiencing all the time, this does not in itself entail that there is a shadow of doubt over whether a subject can *survive* periods of unconsciousness. In some circumstances (e.g. severe brain damage) there might be, but not if the periods in question are bridged by a continuous capacity for continuous consciousness — not if C-relatedness is present and *unimpaired* throughout.

What of uploading, and the prospects of surviving it? Since there is no requirement that C-systems which belong to the same subject must belong to the same physical systems, it is (in principle, at least) perfectly possible for a single subject to migrate from one physical system to another; all that matters is that this subject's capacity for continuous consciousness is not disrupted by the change in material substrate. Since gradual uploads of the sort Chalmers considers — consisting of the progressive nano-engineered replacement of biological cells by silicon-based functional surrogates — do not disrupt capacities for consciousness, they are unambiguously person-preserving. Since destructive uploads clearly *do* disrupt this capacity, at least in cases which involve the original subject being reduced to a passive collection of stored data, they are definitely *not* survivable.

Of course, this is assuming that something along the lines of the C-theory provides us with the correct account of personal identity. According to the (wide) psychological account, destructive uploads *are* survivable, since they lead to the creation of conscious subjects whose psychological systems are both indistinguishable from, and appropriately causally related to, the psychological systems of the original pre-upload subjects. Although I believe the C-theory is superior to the psychological approach in several respects, this is not the place for a defense of this claim; my main concern here has been to bring to light some of the options available to those who take phenomenal consciousness as their guide to personal identity — which proponents of the psychological approach do not. That said, before moving on there is one point to note. It is not unlikely that a sizeable part of the credibility of the psychological approach derives from the assumption that chains of causal dependency linking earlier and later psychological states are the *only* form of mental connection capable of bridging interruptions in consciousness. At an intuitive level it is by no means clear that this sort of connection is sufficient, in and of itself, to constitute personal survival. But if it is this sort of connection which permits us to survive periods of unconsciousness, how can we consistently deny that it suffices for our persistence on other occasions? The availability of the C-theory changes the situation dramatically: interruptions in consciousness can also be bridged by the continuous potentiality for continuous consciousness. If we are essentially subjects of experience — i.e., beings with the *capacity* for consciousness, a capacity which is sometimes exercised, and sometimes not — then this form of continuity is manifestly both necessary and sufficient for our continued existence. In which case, destructive uploads — or

variants of teletransportation — which disrupt or destroy this capacity cannot be survived; at least not by beings such as we.[10]

4. Simulation Multiplication

Inevitably, the precise way in which the C-theory resolves the uncertainties surrounding uploading — in some if not all of its forms — will not be seen as unalloyed good news by everyone. In particular, the thesis that the kind of psychological continuity which is preserved in destructive uploads does not provide for personal survival may seem like bad news for those impatient to upload themselves into virtual (quasi-) paradises. We do not yet have the scanning technology required for destructive uploading, but it is not inconceivable that these might be developed in the foreseeable future, whereas the advances in nanotechnology and neuroscience needed for a gradual upload may seem a far more difficult and distant prospect. But those who are eager to attain a post-human condition need not despair. Although developing the science and technologies required for gradual upload may always be beyond unaugmented human-kind, there is no reason to think a superintelligence will be unequal to the task. If a singularity does occur, taking leave of this world for a Diaspora-style heaven may be a very real possibility — perhaps in the not too distant future.

However, the possibility that future technological developments will make it very easy to create very large numbers of real-seeming virtual worlds gives rise to further issues. These worlds aren't just havens that ordinary humans might move to via uploading, they can impact on the way *all* conscious beings conceive of themselves and their lives. As things stand, our abilities to create and control streams

[10] There is, of course, more to say. Many of us might not *now* be able to view destructive upload as truly survivable, but could this change? Mark Johnston has recently argued that facts about personal identity are determined solely by our own responses to real and imaginary cases of putative survival; by changing these responses we can change what we can survive: 'If refiguring our identity-determining dispositions can open us up to, or close us off from, certain forms of survival, then there is a sense in which our natures are Protean … if we could refigure our identity-determining dispositions then what we are (in the relevant sense) *capable of surviving* would change' (Johnston, 2010, pp. 283–84). Since these identity-determining dispositions are deep-seated, changing them is neither trivial nor easy — Johnston acknowledges that if it is possible at all it will require time, effort, reflection and appropriate metaphysical instruction — but the rewards for success are potentially very great indeed. Since his case depends on an extended argument for the unreality of a self that I find questionable — see Dainton (forthcoming) — I am not wholly persuaded that Johnston is correct in this. Even so, there is undeniably something amusing in the idea of would-be uploaders willingly participating in (what are, in effect) spiritual exercises in order to secure their passage to a digital paradise.

of consciousness are severely limited. Let's suppose that in a post-singularity future this changes, and it becomes possible to create human-type streams of consciousness, of any length, with any desired characteristics, very easily. A perturbing possibility now looms. If streams of consciousness with a character similar to *this* one could be created in the future — cheaply, easily and frequently — might it not be quite likely that this stream does in fact exist in the future? Isn't there a good chance that we are all enjoying artificially generated experiences?

Let's take this a bit more slowly and carefully. Call the succession of streams which jointly compose the consciousness of a single person from birth until death, a *life-stream*. Despite their differences, your life-stream and mine are of a certain general type: early 21st century human. Let us call these 'type-21 streams'. Now suppose that in the future *very* large numbers of type-21 streams will be created, all of which are indistinguishable, qualitatively and subjectively, from real type-21 streams. To be more specific, suppose the total number of type-21 streams which exist after the year 2100 is around ten times greater than the number which existed in the 21st century itself, with each general type of 21st century life being proportionally represented. If these artificial type-21 streams did exist, the consequences would indeed be perturbing. Are you in a position to tell whether your experience is real or artificially generated? No. Or at least, not if all you have to go on is the character of your experience. What are the odds that your experience is occurring when appears to be, in the early 21st century? Only around one in ten. Although it seems to you that you are a normal human being living at the start of the 21st century, the subjects of all the many artificially produced type-21 streams have very similar experiences and beliefs. These subjects are all mistaken, and so might you be, for it is more likely than not that you *are* one of these subjects.

So, if you think it likely that our descendants, whether human or post-human, will develop and use simulation technologies in this sort of way, you should also think it likely that your current experience are the product of these technologies — that your own world is virtual rather than real. Following Bostrom (2003) I will call this line of reasoning the *simulation argument*, though on occasion I will also refer to the *simulation reasoning*, meaning the same thing.[11] Although not everyone will find the possibility that they are living in a simulation

[11] As Bostrom himself uses the term (see his 2003; 2009a; 2009b), the simulation argument purports to show that at least one of three propositions is true. These are (roughly): (1) that most civilizations go extinct before reaching a high level of technological mastery, (2) that

something to be feared or dreaded, at the very least the simulation argument threatens complacent assumptions about the status of our lives, and for this reason I shall sometimes refer to the *simulation menace* or *threat*. Many will no doubt be inclined to dismiss the simulation argument as a mildly diverting but ultimately unthreatening curiosity. For reasons which will emerge *en route*, I think this would be a mistake. After establishing that the simulation argument is one that should be taken seriously, I will move on to consider some of its implications.

5. Simulation Generation

First some terminological clarifications. Henceforth I shall be using 'simulation' in a very broad way: any state or episode of consciousness is to be regarded as simulated if it is produced by non-standard methods in a controlled fashion (the degree of control may vary). Simulated experiences are of course real experiences in their own right, and while a simulated episode of consciousness may be a re-creation of an original non-simulated stretch of conscious life, it need not be. I shall say that a life (or part of a life) is *virtual* rather than *real* if it is entirely composed of simulated experiences; I shall call the subjects of such experiences *simulants*.

Consciousness can be simulated in different ways, and to different degrees. So far as degree or depth of simulation is concerned, we can contrast *complete* with *partial* simulations. The manufactured type-21 streams mentioned above are examples of complete simulations: every part and aspect of experience is being generated by artificial means. In partial simulations, only some parts or aspects of experience are generated by artificial means. A simulation in which a subject is supplied with a wholly virtual environment (which here can be taken to include all forms of bodily experience) but retains their original psychology is one form of partial simulation. But we can also envisage cases in which the effects of the tampering are restricted to the domain of inner experience. Imagine having your psychology (e.g. memories, beliefs, desires, language skills, personality traits, and so on) replaced with a replica of Napoleon's, and then waking up in your own bed and perceiving your environment in the usual way. In what

most technologically advanced civilizations choose not to create large numbers of simulations of their own pasts, (3) that we are almost certainly living in a computer simulation. I will be commenting on this formulation in §6.

follows, unless otherwise stated, we will be concerned with complete rather than partial simulations.[12]

As for the *ways* in which consciousness can be simulated, it is important to distinguish what I will call *neural* (or *N-simulations*) from *software* (or *S-simulations*). N-simulations result from interfering and controlling the neural hardware in the brain that is ordinarily responsible for producing experience. The simulants in N-simulations are ordinary human beings who are vividly hallucinating in a controlled fashion — of course, this will not normally be apparent to the subjects themselves, who will take themselves to be whoever they seem to be in their virtual worlds. In contrast, S-simulations consist of streams of consciousness which are wholly generated by running programs in computers of some kind; the simulants in S-simulations are *Diaspora*-style citizens.[13] Let us start by taking a closer look at the latter.

S-Simulations

On the assumption that mentality is computational in nature, computerized simulations of human brains could generate conscious mental lives that are subjectively indistinguishable from those generated by biological brains. How remote a prospect is this? It is impossible to be certain, but it is probably somewhat less remote than many suppose. If computer technology continues to advance at the rate it has for the past few decades, it will probably not be long before our most powerful computers equal, or exceed, the processing power and information storage capacity of a typical human brain. Estimates of the latter vary a good deal, but it is currently believed to be of the order of 10^{14}–10^{17} operations per second. Present-day laptops are capable of 10^{9} operations per second, and supercomputers can manage 10^{15}. Raw computational grunt by itself does not count for a great deal; we are still a very long way from knowing enough about the structure and functioning of the human brain to simulate their workings computationally. But progress is also being made on this front. Drawing on thousands of detailed scans of rat brains, the Blue Brain project has succeeded in simulating a 10,000 neuron cross-section (along with 10,000,000

[12] More discriminating distinctions can be drawn. For example, it is possible for a subject to lead a virtual life — in the sense here defined — in the real world (some may recall the holo-doctor in Star Trek *Voyager*). This sort of case will not be relevant to what follows, where we shall be concentrating on simulations in which what is 'perceived' is *not* the real world, at least as normally conceived.

[13] To simplify I will assume that only this universe exists. Other universes make a difference. If all logically possible worlds are real, as Modal Realists believe, then each real life is simulated an infinite number of times, creating a highly significant simulation menace.

interneuronal connections) of the rat cortex on a super-computer.[14] The project's leader estimates that a similarly fine-grained simulation of an entire human brain will be possible by 2020. Even if this proves optimistic, it is not too far-fetched to suppose that such a simulation will be possible within two or three decades. There remains, of course, the problem of understanding enough about the relationship between neural activity and experience to reach the stage where we can directly and efficiently control the kinds of experience a computer-simulated brain will produce. This may well be the most difficult problem of all; but even if it should prove to be beyond *our* abilities — as of now this isn't clear — as far as I can see, there is no reason to suppose it will be beyond the capabilities of a superintelligence. If those who believe a singularity will occur in the next few decades are correct, it is quite likely that S-simulations of ordinary human streams of consciousness will be easily produced not long after.

A few such simulations pose no significant threat, but the situation becomes distinctly menacing if they start being produced by the billion or trillion. Such a situation could develop in at least two ways. The ability to produce menacing simulations could become very widespread, e.g. a hundred years from now everyone might own desktop (or handheld) computers easily capable of running them. If several billion computers were to possess this capacity, even if it were utilized only occasionally, menacing simulations would soon exist in disturbingly large numbers. To make matters a little more concrete, consider the popularity of the *The Sims* franchise over the past decade or so, and suppose the simulated inhabitants are all fully conscious — as they might be, in computer games of the future. The 'god games' enjoyed by our descendants could easily prove to be even more popular, generating menacing simulations in vast numbers.[15]

Alternatively, or in parallel, the capability of running large numbers of simulations might be found in the mega-computers of the not too far-distant future. Bostrom provides an illustration of the potential dangers: 'a rough approximation of the computational power of a single planetary-mass computer is 10^{42} operations per second … Such a computer could simulate the entire mental history of humankind (call this an *ancestor-simulation*) in less than 10^{-7} seconds' (Bostrom,

[14] See Markham (2006) for an impressive overview.

[15] In 2010 Electronic Arts celebrated selling 125 million copies of various versions of the game since its launch in 2000. In gauging the potential threat such software might pose, it should also be born in mind that each copy of the game can generate dozens of virtual characters each time it is played, and that the settings need not be contemporary: *The Sims Medieval* was released in March 2011.

2003, pp. 247–8). If our descendants (whether human or machine) were able to run ancestor-simulations using only a small fraction of the computing resources available to them, they might very well do so, quite frequently.[16] If such circumstances were to obtain, the probability that you and I are inhabiting a computer simulation would be high. That said, the effective programming of a super-massive computer of the kind being envisaged here would be well beyond the capabilities of ordinary humans, and may well depend on the availability of superintelligent A++ systems. But again, if a singularity is as likely as many believe, then this is not the hurdle it would otherwise be.

Those who have grown familiar with the claim that the human brain is the most complex object in the known universe may be surprised to discover that it will not be very long before we are able to construct machines of comparable complexity and computational power. But even if this is the case — and I suspect it is — the simulation menace posed by advances in computer technology could well be far less severe than some would have us believe; it may even be non-existent. For S-simulations to constitute a threat they would have to be truly conscious, and it is by no means certain that this is a real possibility. As I noted in §2, there are influential positions on the relationship between the physical and phenomenal which, if correct, would entail that properly programmed computers could generate human-like streams of consciousness. Many versions of orthodox functionalism have this consequence, as does the rather more plausible dualistic (or non-reductive) form of the doctrine sympathetically explored and expounded by Chalmers.[17] But as Chalmers himself would concede, non-reductive functionalism is at best a possible solution to the matter-consciousness problem. Other leading positions in the philosophy of mind — including familiar forms of both dualism and materialism — are resolutely hostile to the notion that consciousness is an organizational invariant. If, as many believe, phenomenal properties are material in nature, then it may very well be that a human-type consciousness requires a human-type brain, or at least a biological system

[16] Ancestor-simulations replicate the *mental* history of humankind, not the entire universe — as Bostrom notes, this would be unfeasible. So although ancestor-simulations reproduce the appearances of a physical world, they do not attempt to simulate all the physical processes within the objects its subjects perceive (e.g. their houses, the interior of the Earth, the distant stars they can see at night).

[17] Or in a little more detail: 'given any system that has conscious experiences, then any system that has the same fine-grained functional organization will have qualitatively identical experiences … we might call [my doctrine] *nonreductive functionalism*. It might be seen as a way of combining functionalism with property dualism' (Chalmers, 1996, pp. 248–9).

of a similar kind.[18] Of course, we cannot be certain of this. We do not know which parts or aspects of the physical processes in our brains are responsible for producing consciousness; consequently, we cannot rule out the possibility that the relevant physical processes could be replicated in very different physical systems — perhaps even silicon chips. But this is no more than a possibility, and in all likelihood, one that is quite remote if materialism is true.

So are S-simulations possible? The situation is clearly far from clear-cut. Whereas functionalists — of both reductive and non-reductive persuasions — have good reasons for being very wary of future developments in computer technology, those who subscribe to different options in the philosophy of mind have far less reason to feel greatly concerned: simulations that are not truly conscious pose no menace whatsoever.

N-Simulations

Those who find all varieties of S-simulation implausible cannot afford to be complacent. For there is a further source of menacing simulations, one that has received less attention, and one that does not depend on adopting controversial positions on the matter-consciousness issue.

Many of us have experienced fully realistic hallucinations, whether drug-induced or in ordinary dreams. Hallucinations of this sort are typically *uncontrolled*: we cannot determine in advance the type of virtual world we will hallucinate, or the role we will play in the scenarios which unfold. As noted earlier, this may very well change, and in a dramatic fashion: post-singularity advances in brain science may make it possible to generate complex and precisely controlled hallucinations — or N-simulations — safely, easily and reliably. Almost inevitably, some of these controlled hallucinations will constitute menacing simulations.

In fact, there are several ways in which this could come about. One way of generating N-simulations would be to use the kinds of neural implant and human-machine integration that are already familiar from science fiction. Interacting with computers mechanically — using keyboards, mice, touch-sensitive screens, etc. — is a cumbersome business, and a good deal of research is going into ways of facilitating

[18] It is difficult to see how experiential states could *be* physical states if the latter only have the range of properties recognized by contemporary physics, but the more credible forms of materialism do not circumscribe the physical in this manner. Lockwood (1989), McGinn (1999), Searle (1992), Strawson (1994) all defend version positions which combine materialism with a realist (non-reductive) view of consciousness.

the process. Among the methods already being considered, at least by the more adventurous researchers, are direct brain-computer interfaces. At present, such techniques are at a fairly primitive stage of development, but this will no doubt change.[19] A hundred years from now, children could be growing up with implants buried deep in their heads, implants that both track and keep pace with their neural development, and allow their minds to interact directly with computers, on a number of levels, in a variety of ways.

It is not difficult to envisage some of the uses to which this sort of interface might be put. Your thoughts could be transmitted directly into someone else's mind — provided you were both hooked up to the same computer network. Forgetfulness would be largely a thing of the past: your thoughts and experiences could easily be backed-up on a computer file, ready to be called on when required. More relevant to our purposes, fully immersive virtual reality would also be a possibility. There will be no need for you to wear a suit and visor to interact with machine-generated virtual worlds, your implants will perform the necessary tasks. Your sensory experience will be directly machine-controlled, via stimulation of the appropriate areas of the sensory cortex. The movements of your (simulated) body through virtual environments will be under your control, but there will be no need for you to actually move your physical body: the ways you *intend* to move your body will be detected by implants in your motor cortex and elsewhere, and this information will be used to generate corresponding movements of your virtual body. It will be possible to have a fully realistic experience of (say) flying a plane through narrow mountain passes while remaining motionless on a couch. You might even *believe* yourself to be an experienced pilot: your implants could ensure that a suitable set of false memories temporarily override your real memories. Alternatively, you might believe yourself to be an ordinary 21st century person, leading a typical life in a (virtual) 21st century environment.

A materialist might argue that brain-computer connections required for this sort of simulation would be so invasive and pervasive that their presence inside a brain would be incompatible with the

[19] Indeed, brain-computer interfaces have already left the laboratory and entered (some) homes: the Emotiv EPOC is a headset which allows gamers to interact directly with their consoles; in the company's own words, the system 'is a high resolution, neuro-signal acquisition and processing wireless neuroheadset. It uses a set of sensors to tune into electric signals produced by the brain to detect player thoughts, feelings and expressions and connects wirelessly to most PCs' — all this for only $299! See http://www.emotiv.com For an indication of the (quite impressive) level of sophistication of current computer-based 'mind-reading' techniques, see Naselaris *et al.* (2009).

production of conscious experience. Given our ignorance of the physical processes underlying consciousness, this possibility cannot be ruled out, but there no reason to suppose it very likely. After all, the envisaged interfaces would not replace neurons as experience-producers, they would merely provide ways of artificially controlling the triggering of neurons, or neural circuits — and we already know this to be possible on a small scale. True, the required nano-scale technology is far beyond anything we are capable of producing at present, and even if it were not, our understanding of the brain's functioning is not sufficiently advanced for it to be deployed effectively. But anyone inclined to think this will continue to be the case should bear in mind two considerations: first, the difference the availability of (helpful and cooperative) superintelligent computers would make, and second, Arthur C. Clarke's dictum that any sufficiently advanced technology is indistinguishable from magic.

If the envisaged interface technology were to become commonplace, then given time, N-simulations could easily be produced in sufficient numbers so as to become menacing. People might take virtual reality 'trips' to the past quite frequently. They would certainly be used on an occasional basis during history lessons, and more intensively by historians, amateur and professional, with a particular interest in what it was like to live during certain periods of the past. But such trips might also be taken — far more frequently — for entertainment purposes. The soap operas of the future might well have an immersive/interactive character their present-day counterparts lack, computer games likewise. Already there is evidence that being able to enter and explore virtual worlds is likely to prove extremely popular. Over the past few years the numbers of people participating in MMPORGs ('massively multiplayer online role-playing games') has expanded dramatically — the currently dominant *World of Warcraft* currently has close to 12 million participants.[20] The addictive properties of these virtual worlds is well known; it is not uncommon for gamers to absent themselves from this world for hours per day. With the advent of fully immersive virtual reality technology, pastimes of this sort will probably prove to be still more popular. As is easily shown, the numbers soon add up.

Our descendants may 'visit' the past quite frequently, but since few are likely to want to spend significant portions of their lives in N-simulations, the concept of a *life-stream* introduced earlier is no longer

[20] According to a Blizzard press release, see http://us.blizzard.com/en-us/company/press/pressreleases.html?101007

appropriate as a basic unit of simulated consciousness. Something of briefer duration is required. So, for present purposes, let us take *day-long* streams of uninterrupted consciousness — *D-streams* for short — as our working units. (An even shorter unit could be selected, but as will become evident, the upshot would not be greatly different.) We shall take as our class of *menacing* D-streams those simulated streams that resemble the sort of experiences enjoyed by actual inhabitants of the year 2011 — call these *MD-streams*.

Assuming the current population of the Earth to be seven billion, there are just over 2.5×10^{12} D-streams for the year 2011. If a similar number of MD-streams exist in the future, the odds of the experiences you are currently having being simulated rather than original look to be around fifty per cent. Should the numbers of MD-streams created in the future be greater, your chances of living among the original inhabitants of the year 2011 will be correspondingly smaller.

In fact, the number of MD-streams created in the future could easily be far higher. Call the time at which N-simulations become commonplace occurrences the *C-threshold*. Let us suppose that from the C-threshold on, every future human being takes one virtual reality trip to the year 2011 during their lifetime, and that these trips are varied in character. If we now suppose that human civilization lasts for ten thousand generations after the C-threshold, and has an average population of ten billion, there will be 1.0×10^{14} MD-streams, compared with 2.5×10^{12} original D-streams. With forty simulated streams for every real stream, you have a one in forty chance of actually being alive in the year 2011.[21]

It is not only materialists who should be open to the possibility of N-simulations, dualists should be too. Even if our experiences unfold within immaterial substances, it is evident that our minds are profoundly dependent upon our brains. No contemporary dualist would be inclined to deny that the course of our sensory experience is dependent upon the neural activity within our brains, and this fact alone opens up the possibility of controlled hallucinations of a limited kind. But dualists should also recognize that appropriate neural manipulation could impact upon our conscious beliefs, intentions and desires. Intoxicants do not merely make it harder to control our bodily

[21] On more optimistic scenarios, your predicament is even more precarious. If humankind has a long history — one million generations exist after the C-threshold, say, with constant or improving technology — and a larger average population during this period (a hundred billion, say) then we can expect a total of around 1.0×10^{17} MD-streams to occur, which would reduce your chances of being alive in 2011 to around one in fifty thousand! In this case, even if only one in a thousand people ever take a virtual reality trip back to 2011, the chances that you are really living in 2011 are still only one in fifty.

movements, they make it harder to *think* clearly, and there are numerous forms of brain damage that have more far-reaching (and often permanent) effects on our personalities and cognitive functioning, memory included. If brain damage can result in the permanent loss of certain memories, is it not likely that the memories to which we have conscious access depend on information stored in our brains? In which case, appropriate neural manipulation could lead a 23rd century person to have access to apparent-memories of the sort a 21st century person would have had.

But there is a further point to note, one that is relevant to materialists as well as dualists. Brain-computer interfaces of the kind I have been considering offer the possibility of very tightly controlled hallucinations, but there are undoubtedly other ways of inducing similarly life-like N-simulations, even if they offer rather less potential for fine-grained control. Ordinary, unaugmented, human minds are able to fashion richly-detailed and real-seeming virtual realities all on their own, almost effortlessly. Ordinary dreams provide evidence both of this, and our ability to spin complex virtual worlds from limited and/or fragmentary evidence. I expect most of us have found ourselves having vivid dreams set in (say) the 17th century shortly after watching a film set in the same period. Although the dreamed environment in such cases is inspired by what was seen onscreen, it often has a depth and complexity all of its own.[22] Future methods of experience-induction could easily exploit these ordinary abilities. All that would be required is a safe and reliable drug which enabled people to enter a dream-like state at will, and also direct the general direction of the subsequent (fully life-like) hallucination — the framework for the latter could be supplied by a little prior reading, or the watching of video footage (e.g., of a 21st century televised soap opera). This method of controlling hallucinations could be put to the same uses (in, say, education and entertainment) as the computer-driven variant we considered earlier, and so is likely to be widely employed. So far as I can see, this method of inducing (partially) controlled hallucinations is not ruled out by any philosophical conception of the mind. It is also quite likely to prove attainable, perhaps quite soon — and may well not even require the advent of superintelligence.

[22] Might *ordinary* dreams constitute menacing simulations? Perhaps, but I am inclined to think not, simply because I suspect my dream-experiences are somewhat less vivid and more course-grained than my ordinary waking experiences. I am not alone in this (see Flanagan, 2000, pp. 173–4). Of course, we cannot rule out the possibility that we are living in the dream-worlds of beings whose waking consciousness is far richer than our own.

6. An Issue of Principle: Is the Reasoning Self-Defeating?

Unlike those familiar forms of sceptical reasoning that invoke powerful deception-mongering demons and suchlike, the simulation reasoning rests on more mundane considerations: more or less plausible projections of current technological and social trends. This novelty grants the simulation reasoning an uncommon force, but it also leaves it vulnerable to an objection along these lines.

> The simulation argument relies on certain empirical premises concerning how our world is likely to turn out. If we come to accept that it is likely that we are living in simulations, we surely no longer have reason to accept the relevant empirical premises. The simulation reasoning is self-defeating, and can therefore safely be dismissed.

More generally, it might seem that any argument rooted in empirical considerations that yields a radically sceptical conclusion is inherently unstable. No sooner is the sceptical conclusion generated the supporting empirical considerations are blasted away and the conclusion collapses.

Bostrom has recently responded to this objection in his 'Simulation Argument FAQ' (2009b). While his argument is useful as far as it goes, in some respects it does not go far enough.

As Bostrom formulates it, the simulation argument is supposed to show that at least one of the following claims is true: (1) the human species is very likely to go extinct before reaching a technologically advanced posthuman stage; (2) it is very unlikely that any posthuman civilization will run large numbers of simulations of their own history, (3) we are almost certainly living in a computer simulation. Now, in his response Bostrom maintains that it is wrong to claim that we can't have *any* reliable information about the underlying external reality if we actually are in a simulation. For even if we are simulants we can know the following two conditional claims: (a) if we are in fact living in a simulation, then the underlying reality contains at least one simulation (this one), or (b) if we are not in a simulation, then there is no reason to doubt our empirical beliefs (e.g. to the effect that a technologically advanced civilization will have the ability to create vast numbers of simulations). If (a) is true, and we are in a simulation, then we know that (3) is true. If (b) is true, then we are not living in a simulation, and so we know that (3) is false, but (1) or (2) may still be true. So even if I am simulated I can know that either (a) or (b) is true, and hence know that at least one of (1), (2) and (3) is also true.

Formally speaking, this response is enough to establish that Bostrom's formulation of the simulation argument does not succumb

to the charge that it is self-defeating. However, this line of response also diminishes the interest of the simulation reasoning. If we follow Bostrom's line the simulation argument is no longer an empirically grounded *personal* threat — the reasoning no longer takes us (as individuals) from the empirical premises to the conclusion that we ourselves are simulants. For the argument in the latter guise to work it must be possible for me to *combine*, consistently and coherently, the belief that I am living in a simulation with the belief that most of my empirical beliefs are true. Bostrom's response does not allow me (or you, or anyone else) to be justified in holding both of these beliefs. As he presents our predicament, if I am in a simulation then the only well-founded belief that I am entitled to with regard to the external world is that it contains at least one computer simulation, the one I inhabit. If this is *all* I am entitled to know, then my original empirically based beliefs for thinking that I am a simulant are no longer well-founded. The hypothesis that I'm living in a simulation has, in effect, been reduced to the bare possibility of traditional scepticism. And of course, what goes for me, goes for you.

But we are not yet done, for the epistemic instability charge can be countered in another way. As a first step, consider these two strongly contrasting simulation hypotheses.

> SH1 Simulated lives vastly outnumber real lives; the environments simulated subjects inhabit are invariably utterly unlike the real world.
>
> SH2 Simulated lives vastly outnumber real lives; without exception, the environments simulated subjects inhabit are almost exact replications of the real world.

I shall call simulations *faithful* if they resemble some part or period of the real world quite closely in most respects, and *unfaithful* if they do not. SH1 presents us with an extreme example of an unfaithful simulation, SH2 is an extreme example of a faithful simulation. Now, considered purely as a prediction as to the quantity and character of simulations likely to exist, SH1 is unproblematic. It quickly becomes problematic if it is incorporated into a simulation argument thus: 'I have empirical grounds for believing that (i) simulated lives outnumber real lives, (ii) simulated lives are invariably unfaithful, (iii) it is highly likely that I am myself a simulant.' The epistemic predicament of anyone subscribing to this combination of claims is precarious. No one who subscribes to both (ii) and (iii) can reasonably claim to have firm *empirical* grounds for believing (i). If you believe you inhabit a simulated world that is drastically misleading with respect to what is really going on, then the fact that this world has features which make it

plausible to believe simulations will be created *en masse* does not provide you with a reason for supposing that these simulations really will be created — and you therefore have no reason to conclude that it is highly likely that you are yourself a simulant, or at least, not on empirical grounds.

The situation with regard to SH2 is very different. If I have good empirical grounds for believing that nearly all simulations that will ever be produced will be faithful, there is no tension whatsoever in my *also* subscribing to (i), and hence thinking it likely that I am myself a simulant. In these circumstances, even though I believe it is highly likely that I am living in a simulation, I also have good reasons (or so we are supposing) for believing that the world I am acquainted with in my simulation accurately depicts how things really are outside my simulation. Consequently, if the socio-technological trends in my (very likely simulated) world are such that it seems probable that large numbers of simulations will be created, it is perfectly reasonable for me to conclude that this will occur in the non-simulated world — and so it is perfectly reasonable for me to retain my empirically-grounded belief that I am myself likely to be a simulant.

So there are conditions under which the simulation argument is *not* self-undermining. Of course, SH1 and SH2 are extreme and rather implausible simulation hypotheses. But although more plausible hypotheses bring further complications, they do not change the overall picture: even if there are conditions under which the simulation argument is self-undermining, this is by no means invariably the case. To illustrate, suppose you have highly compelling reasons, grounded in detailed analyses of technological and social trends, for subscribing to this simulation hypothesis:

> SH3 Simulated lives outnumber real lives by about ten to one; my type of life has as much chance of being simulated as anybody else's; roughly half of all simulations are faithful, half are unfaithful.

Even if your type of life does has as much chance of being simulated as any other, it would be a mistake in this case to conclude that the odds of your own life's being real are of the order of one in ten. When calculating these odds all *un*faithful simulations must be discounted. For as we have just seen, although it is perfectly consistent to believe on empirical grounds that such simulations exist, it is not consistent to take the further step and infer that their existence makes it more probable than it would otherwise be that you are yourself a simulant. But as we have also seen, there is no such difficulty in the case of faithful simulations. If you have good reasons for believing that simulations

outnumber real lives by around ten to one, but only fifty per cent of simulations are likely to be faithful, it is reasonable for you to conclude that that the odds of your life being real are around one in five.

We can draw a more general lesson from this. The seriousness of the simulation menace in any particular context depends on the degree to which the simulation hypothesis is credible, and the predicted ratio of faithful to unfaithful simulations. Other things being equal, the greater the proportion of unfaithful simulations, the lesser the simulation menace, the greater the proportion of faithful simulations, the greater the simulation menace.

A further complication concerns the property of faithfulness itself. To be menacing a simulation need not accurately reproduce some portion of the real world in every respect. All that matters is accuracy in those respects that are relevant to assessing the likely quantities and types of simulation that a given world is likely to produce over a given period. Call this *simulation-relevant* (SR) *accuracy*. For a simulation to be SR-accurate it will conform to (apparent) natural laws similar to those that obtain in the real world, and the broad sweep of social and technological trends during the relevant period — in our case, the early 21st century — will also be similar to those in their non-simulated counterparts. SR-accuracy does not require fidelity in matters of historical and biographical detail. To make matters more concrete, think of works of fiction. Nearly all contemporary novels — with the notable exception of most fantasy and some science fiction — are highly SR-accurate despite the fact that most of the characters and events described in such fictions do not exist. To put it another way, if a character in a typical novel with an early 21st century setting were to attempt to assess the simulation menace in their world, the relevant data available to them would be much the same as is available to you or I.

Taken together these considerations have a threefold impact. Since judgments as to SR-accuracy will be based on estimates of a number of difficult-to-estimate variables, in many (but not all) cases it will be impossible to arrive at precise numerical assessments of the simulation menace. However, for the menace to be real we do not need anything particularly fine-grained: ‘negligible’, ‘moderate’ and ‘high’ will serve nicely. Secondly, if only simulations scoring highly on SR-accuracy are menacing at all, the overall simulation menace may well be somewhat reduced, at least for the more moderate brands of simulation hypothesis. Those who think it likely that the future will contain immoderate number of large-scale S-simulations probably do not need to lower their estimates significantly — Bostrom’s ancestor-

simulations, for example, all score very highly in SR-accuracy.[23] By contrast, those of us who find these scenarios implausible, and who lay more weight on the likely existence of many small-scale N-simulations, many of which may not be SR-accurate, may confidently downgrade our assessments of the simulation menace. Downgrade but (probably) not dismiss: it seems plausible (to me at least) that a good proportion of future simulations will be SR-accurate.

Finally, and importantly, we have uncovered an important constraint, of the transcendental variety, on the *type* of simulation that it is coherent to think one may exist within on empirical grounds. If you are led by the simulation reasoning to the conclusion that there is a fair probability that you are inhabiting a simulation, you have every reason to suppose that the non-simulated world is not *too* dissimilar to your world — in effect, you may be living in a total fiction, but you are not living in a total fantasy. While it remains perfectly possible that you are living in an environment that bears no resemblance to how things really are, coming to accept the simulation reasoning, in itself, should not lead you to think it more likely that this is so.

This constraint may do something to lessen the anxiety-provoking consequences of the simulation argument, but I think it fair to say that it by no means eradicates them entirely. The thought that one is inhabiting the equivalent of a work of fiction is disturbing enough!

7. Some (Further) Varieties of Virtual Life

The distinction between N-simulations and S-simulations reflects one way in which simulated lives can be subjectively indistinguishable but different in kind. There are other ways, and it will be helpful to have a few of these in view before proceeding further.

Active v. Passive (A-simulations v. P-simulations) The subjects of A-simulations are confined to virtual environments, but in all other respects they are free agents — or as free as any agent can be. Their actions are not dictated by the virtual-reality program, they flow from their own individual psychologies, even if these are machine-implemented. A P-simulation, by contrast, is a completely pre-programmed course of experiences. The subjects of P-simulations may have the impression that they are autonomous individuals making free choices, but unlike their A-simulation counterparts, they are deluded: all their conscious decisions are determined in advance by the virtual reality program. Such subjects have *apparent* psychologies — their consciousness is subjectively similar to that of someone with an active

[23] I suspect this is why Bostrom seems little concerned by the allegation of epistemic instability.

psychological system, so they have apparent memories, hopes, fears, etc. — but their real psychologies are entirely suppressed (or even non-existent).

Original Psychology v. Replacement Psychology (simulations$_{OP}$ v. simulations$_{RP}$) In A-simulations, a 'replacement psychology' is an artificially-generated system of beliefs, desires, memories, intentions, preferences, personality traits and so forth that supplants a subject's own ('original') psychology. The same applies in P-simulations, the difference being that the replacement psychology is only *apparent*, in the sense just introduced. There is a sense in which the inhabitants of simulations$_{RP}$ are doubly deceived: not only is their environment not what it seems, neither are their minds.

Communal v. Individual (C-simulations v. I-simulations) A C-simulation is a virtual environment shared by a number of different subjects, each possessing their own distinctive individual psychology (even if these are machine-implemented). An I-simulation is restricted to a single subject. Of course, the subject of an I-simulation may meet what they take to be other people in their virtual worlds, but these 'others' do not possess their own individual autonomous psychological systems — they are not subjects in their own right, merely parts of a machine-generated virtual environment.

These options can be combined in various ways, e.g., a simulation of type AC$_{RP}$ is active, communal with replacement psychology, whereas a simulation of type PI$_{OP}$ is passive, individual with original psychology. There is a total of eight permutations:

AI$_{OP}$: Active/Individual/Original Psychology
AI$_{RP}$:Active/Individual/Replacement Psychology

AC$_{OP}$:Active/Communal/Original Psychology
AC$_{RP}$:Active/Communal/Replacement Psychology

PI$_{OP}$:Passive/Individual/Original Psychology
PI$_{RP}$:Passive/Individual/Replacement Psychology

PC$_{OP}$:Passive/Communal/Original Psychology
PC$_{RP}$:Passive/Communal/Replacement Psychology

Assuming that each of these modes could be generated by either N-methods or S-methods, we have a grand total of sixteen distinct kinds of (subjectively indistinguishable) virtual life. But the situation may not be quite so complex. A strong case can be made for thinking that a truly *communal* simulation of the passive variety is impossible. There is nothing impossible in the idea of a number of subjects simultaneously playing out roles in similar and coordinated hallucinations, but unless these subjects can interact and converse with one another they can scarcely be said to constitute a genuine community, and this

cannot really occur in P-simulations (although this might not be obvious to the simulants concerned). For this reason it seems right to regard all P-simulations to be of the individual variety. This brings our grand total down to twelve.[24]

8. A New Scepticism – Or A New Metaphysics?

Ancient Greek sceptics argued that since our senses can deceive us we can never be justified in supposing that the world is how it seems, but the idea that there might not even *be* an external world never occurred to them.[25] For the latter hypothesis to be thinkable consciousness must be construed as a self-contained and potentially autonomous realm of existence in its own right. Descartes was the first to articulate this conception clearly, and drew the (now) obvious sceptical conclusion: our experience could be (subjectively) just as it is even if the reality external to our consciousness is very different from how we believe it to be on the basis of our experience. Consequently, we cannot be certain that the physical world exists.

There are similarities as well as differences between Simulation Scepticism (as we might as well call it) and Cartesian Scepticism. So far as the existence of a mind-independent reality is concerned, Simulation Sceptics are as one with their Greek predecessors: it exists (it is where the simulations are being run). But Simulation Sceptics are as one with Cartesian Sceptics when it comes to the status of our current consciousness: both hold that it could be wholly virtual, a detailed and convincing hallucination. However, for the Cartesian this conclusion relies on the world external to our minds being very different from how it appears (e.g., reality might consist of nothing but your consciousness and a malicious Demon). Simulation Sceptics, by contrast, derive their conclusion from the assumption that our experience is a broadly reliable (SR-accurate) guide to the character of that portion of the external reality it seems to concern.

Simulation Scepticism is in this respect less radical than its Cartesian counterpart, but it is also less of a blind alley. Different

[24] To simplify, I overlook here the fact that the distinction between N-simulations and S-simulations may not be absolute: the consciousness of future humans (or post-humans) may be sustained by a combination of neural and artificial means, and neurons themselves may be genetically manipulated. I also ignore the fact that in some logically possible worlds, simulations are created by quite different means (e.g. magic). It should also be noted that simulants of different types can coexist, e.g., *The Matrix* films feature a combination of N-simulations (ordinary humans) and S-simulations (the 'agents'), both active, coexisting in a single C-type virtual environment.

[25] Or so it has been argued (*cf.* Burnyeat, 1982).

hypotheses as to how the future may turn out, or what the universe may contain, render the hypothesis that we are leading virtual lives more or less likely, and these hypotheses can be refined, explored and evaluated. Unfortunately, this gain comes at a price. The threat posed by Simulation Scepticism is far more *real* than that posed by its predecessors. Cartesian scepticism is hard to refute, but as Hume noted, it is also hard to take seriously. Few of us spend much time worrying about the possibility that reality could be radically different from how it seems. Simulation Scepticism reveals that even if reality *is* largely as we believe it to be, there could be a high probability that our actual condition is very different from our apparent condition. As things stand, with simulation technology still at a primitive level, many will find Simulation Scepticism as hard to take seriously as its Cartesian counterpart. This will change as the technology advances.

However, it could be a mistake to regard the simulation argument as leading to a sceptical hypothesis of even a modest kind. An alternative approach is to construe the reasoning as leading to a novel *metaphysical* hypothesis concerning the underlying nature of one's environment. For some simulation scenarios, if not all, there is a good deal to be said for adopting this line. Suppose, for example, that you come to believe that you are living in a large-scale, long-lasting, communal simulation. You may initially be inclined to think: 'The material world of which I took myself to be an inhabitant does not exist — all my experience has been delusory, akin to a dream or hallucination.' This is an overreaction, albeit an understandable one. Your world may not be 'material' in the usual sense of the term, but there is certainly a sense in which it is real despite its underlying computational nature. Simulated worlds can be immensely complex, and nomologically speaking, just as well (or badly) behaved as their non-virtual counterparts. Moreover, by virtue of inhabiting a communal simulation, you are not alone. In the company of your fellow C-simulants you are at liberty to conduct empirical explorations of your environment, and agree and disagree on your findings, in all the ways available to the inhabitants of non-virtual worlds. As a consequence distinctions between appearance and reality, between subjective and objective, are as well-founded in your virtual world as they are in any world. Even if your perceptual experience does not directly reveal the real in the way you once naively supposed, it nonetheless reveals a world which possesses many of the defining properties of 'reality': your world is certainly objective, and independent of your mind. Thus construed, the simulation argument does not threaten to undermine our ordinary empirical beliefs — these remain mostly true — what is threatened,

rather, is a doctrine concerning the underlying real nature of the world we inhabit. The import of the simulation reasoning is primarily metaphysical rather than epistemological.

That many simulation scenarios should be construed in this manner has been forcefully argued by David Chalmers:

> …the Matrix Hypothesis is not a skeptical hypothesis. If I accept it, I should not infer that the external world does not exist, nor that I have no body, nor that there are no tables, chairs, and bodies, nor that I am not in Tucson. Rather, I should infer that the physical world is constituted by computations beneath the microphysical level. There are still tables, chairs and bodies: these are made up fundamentally of bits and of whatever constitutes these bits. (Chalmers, 2005, §5)

Chalmers' principle example is a Matrix-type scenarios — a simulation of the N-type AC variety, in the terminology introduced in §7 above — but, as he makes clear, he believes that the metaphysical interpretation extends to S-type simulations. He goes further, and suggests that the metaphysical interpretation can be extended to small-scale, short-lived simulations. In such cases, simulants can still reasonably be regarded as being in perceptual and cognitive contact with a genuine world, it is just that the world is rather smaller than it seems, and as a result, fewer empirical beliefs are true than would be the case for full-scale simulations.

If Chalmers' is correct, the simulation scenarios provide a new twist, and a new clarity, to the Kantian thesis that our world may only be 'empirically real'. As for whether Chalmers' *is* correct, I find much of what he says very plausible. But I do have one worry.

A compelling case can certainly be made for holding that the environments in which the relevant simulants find themselves can legitimately be classed as *worlds* — or 'external realities' — despite their being virtual in nature. However, we can accept this much without also accepting that an external world of this sort constitutes a properly *spatial* world. This is of some significance, because many would also incline to the view that only worlds which are spatial in nature are candidates for being regarded as *physical*. Now, Chalmers maintains that the relevant virtual worlds are both spatial and physical in nature, but it is certainly not obvious that this is the case. Indeed, he anticipates someone objecting to his position along precisely these lines: 'one could suggest that the problem with the matrix is that its spatial properties are all wrong. We believe that external entities are arranged in a certain spatial pattern, but no such spatial pattern exists inside the computer' (Chalmers, 2005, note 14). Chalmers' main line of reply to this objection involves a hypothesis, and two principles. The

hypothesis is that it is at least possible that the microphysical processes throughout our space-time are in fact constituted by computational processes. The relevant principles are these: (1) any abstract computation that could be used to simulate physical space-time is such that it *could* turn out to underlie real physical processes, and (2) given an abstract computation that *could* turn out to underlie physical processes, the precise way in which it is implemented is irrelevant to whether it *does* underlie physical processes (*ibid.*, §5).

Taken together these ingredients deliver the result Chalmers wants, and (here at least) I do not want to question the hypothesis that it is at least possible that the physical world is constituted by computational processes. However, the claim that the way a computation is implemented in the real world can make no difference to whether or not the implementation can legitimately be taken to constitute a genuinely spatial universe, as encapsulated in (2), is more questionable. Why? Simply because some computational implementations are more intrinsically spatial in nature than others.

To illustrate, let's focus on a very simple universe: a finite three-dimensional space, discrete rather than continuous, whose contents comprise a few million particles, which are moving around and interacting in accord with a small collection of simple dynamical laws. As it happens, this universe can most easily and efficiently be modeled by treating it as a three-dimensional cellular automaton, with a small collection of local rules governing the behavior of individual cells (as in Conway's well-known Game of Life, cells change state in response to the cells in their immediate neighbourhood). Now compare two ways in which this abstract computational model could be implemented: (A) the program is run on an ordinary (classical) desktop computer, (B) the program is run on a specially created physical realization of the relevant cellular automaton, i.e., a physical system consisting of a spatially extended three-dimensional grid, whose cells vary in their physical properties in accord with automaton's rules. Is either of these implementations successful in constituting a genuine spatial system? Are the worlds that are created by these computational processes spatial in nature? To deny that implementation (B) generates or constitutes a spatial world would clearly be absurd. But situation is far less clear-cut in the case of implementation (A). The desktop computer in question will be rapidly shuffling patterns of bits through its central processing cores, temporarily storing information in various parts of its RAM, encoding other information in a spatially scattered way on its hard drives, and so forth. This scattered and discontinuous computational process successfully *represents* a continuous space, but the

process itself does not consist of a continuous spatial manifold or medium — this is in sharp contrast to the process in (B), which clearly and unambiguously does. Given this, the claim that in (B) the computational process constitutes a real space, but in (A) the space created is merely a virtual one — nothing more than an appearence — has a good deal of plausibility.[26] As a consequence Chalmers' claim that the precise way in which a program is implemented is irrelevant to whether or not it constitutes or underlies a physical process is itself undermined.

Of course, we have been considering a small toy universe, but so far as I can see, the key point — that not all computational implementations are the same when it comes to their intrinsic spatial properties — still applies on the broader stage. Indeed, some of those who have argued for the possibility that our own universe is a computer, e.g. Zuse (1970) and Wolfram (2002), seem to have also held that at the fundamental level our universe is something akin to a giant cellular automaton (one with a very fine-grained grid). If this hypothesis is correct, then our universe is both spatial and computational in nature. Whether or not the same applies to any virtual worlds being sustained by other (smaller) computers may well depend on their specific mode of implementation.[27]

---

[26] To make matters a little more precise we might say the following. Suppose P is a program for a virtual world. For the space S created by a particular implementation of P to be a genuine rather than merely virtual space, the spatial relations between the regions and material contents of S must closely match those of corresponding computational processes (i.e., the processes which underlie or constitute these regions and contents). So, for example, if region R1 in S is entirely encloses region R2, the computational processes responsible for R1 will entirely enclose those responsible for R2; if R2 occurs between R1 and R3, then the same will apply to the corresponding computational processes. Since 'close match' can be interpreted more or less stringently — e.g. matching topological features might suffice, or we might insist on similar metrical features — the criterion is a flexible one. That said, it is difficult to see that how even the most lax interpretation of 'close' would allow the world created by implementation (A) to count as genuinely spatial.

[27] Chalmers also argues (also in *Matrix*, note 14) that anyone who insists that the implementing level must itself have an appropriate spatial structure before it can be counted as constituting a physical world is running counter to the spirit of contemporary physics, where the claim that our own physical space is not a fundamental feature of the world, but rather an emergent one, is being taken seriously by leading physicists and cosmologists. Physicists are indeed taking this claim seriously, but as far as I can see, in several well-regarded approaches (e.g. loop quantum gravity) all that is being dispensed with is space construed as an entirely autonomous or independent background medium; spatially related material particulars remain very much in the frame. The holographic principle of 't Hooft and Susskind *could* be interpreted as entailing that our universe is really two dimensional (Susskind, 1994, §1: 'Instead of a three dimensional lattice, a full description of nature requires only a two dimensional lattice at the spatial boundaries of the world. In a certain sense the world is two dimensional and not three dimensional as previously supposed.')

9. Simulation Ethics

So much for how we should conceive of virtual worlds. What of how we should act in them? Would taking the simulation argument seriously have any practical or ethical implications for how we should lead our lives?

If we knew what kind of simulation we were living in, the answer would clearly be in the affirmative. In an I-simulation what appear to be other people are no more than the appearances of such. Simulants such as these have the same ethical status as characters in contemporary computer games, and they could be treated accordingly — though of course, *un*like characters in contemporary computer games, they can hit back, and so some caution is in order. The knowledge that one might be living in an I-simulation can also console. Anyone who has had cause to regret an action because of its consequences — and that includes most of us — will be cheered by the thought that these consequences might not in fact be real. The downside of this is that the same would apply to one's most valued relationships. The knowledge that one is living in a passive simulation, and so cannot be held responsible for one's actions or omissions, brings similar advantages and disadvantages. Passive simulants cannot be blamed for their mistakes or wrongdoings, but neither do they merit admiration for their achievements.

Further examples could be supplied, but the point is clear. Simulation scenarios have practical and ethical consequences, and these consequences vary depending on the type of simulation involved. And there lies the problem. Even if we knew the precise probability of our lives being simulated, which we don't, this knowledge would be useless for most practical purposes unless we also knew the sort of simulation being run, along with and the intentions and preferences of the simulators. Unless and until such knowledge is forthcoming, it is probably best to continue much as we would otherwise do.[28]

This way of interpreting this principle is controversial, but even so: there is a big difference between a world of two spatial dimensions and a world of *no* spatial dimensions!

[28] The difficulty of attempting to second-guess the likely preferences of simulators is illustrated by Hanson's (2001) recommendations as to how one should act so as to reduce the risk of the curtains being brought down on one's virtual world. He concludes thus: 'If you might be living in a simulation then all else equal it seems that you should care less about others, live more for today, make your world look likely to become eventually rich, expect to and try to participate in pivotal events, be entertaining and praiseworthy, and keep the famous people around you happy and interested in you.' I am not sure that anyone could conform to *all* these injunctions simultaneously, but even attempting to do so might well make one so obnoxious as to hasten one's end.

However, even if this approach is optimal for many, it may not be appropriate for everyone. Anyone who thinks they are living an especially interesting life, and so having experiences which future simulators might be particularly interested in 'sharing', will be led to the conclusion that their life has a greater than average chance of being virtual. It is hard to predict what effects this realization might have on (say) political and military leaders of the future, but they may well not be wholly beneficial, to put it mildly.

There is a second ethical issue to consider. Since simulated lives are subjectively indistinguishable from the real thing, their creation is by no means a trifling matter, morally speaking. Even if our descendants (whether human or machine) develop the means of producing such simulations easily and cheaply, might they choose not to do so? Might ethical scruples eliminate or at least diminish the threat posed by the simulation argument? Since we are dealing with the future it is impossible to be sure, but there are certainly some reasons for thinking it unlikely.

One of the uncertainties derives from the fact that future simulation technologies, or at least a significant proportion of them, may well be in the hands of superintelligent machines. Although we might try to ensure that these AIs share our values, as Chalmers notes in *Singularity* §6, there is no guarantee that we will be successful. If a superintelligence emerges through a digital version of natural selection — a not unlikely eventuality, in my view — our influence on it will inevitably be limited: we will probably have a very imperfect understanding of its program. But even if we do design an AI+ system, and choose which values to instill into it, there is no guarantee that the more sophisticated A++ designs which follow will share these values; after all, these systems will be far more intelligent that we. Although Kantians are of the view that rationality and morality are by their natures inseparable, a compelling proof of this connection has proved elusive. Pulling these points together: when it comes predicting the value-systems which future AIs will subscribe to, there is very little we can be sure of. As a consequence, even if there are ethical reasons for not creating virtual lives which *we* (or our human descendants) find compelling, we cannot be sure that A++ systems will share this view.

There are further complications. It is easy to conceive of simulations which most of us would judge to be morally abhorrent, and thus wrong to create. An obvious example would be S-simulations of entire virtual worlds all of whose inhabitants suffer nothing but perpetual pointless torment. It is far from inconceivable that our

descendants (human or machine) will forbear from creating such things. That said, even this is by no means certain. The plot of Iain M. Banks' recent novel *Surface Detail* (2010) revolves around virtual hells, deliberately created and maintained as fit punishments for the wrong-doers uploaded into them; the various gruesome tortures inflicted on the unfortunate inhabitants would soon put an end to any flesh and blood human, but can be sustained for seeming-eternities in a virtual world. Can we be sure that none of our descendants will find this mode of punishment appealing?

In any event, not all large-scale simulations are clearly morally abhorrent, far from it. Would creating an ancestor-simulation — a complete S-simulation of human history up until the present time — be morally wrong? It is not at all obvious that it would. The sum total of human misery may be immense, but so too is the sum total of human happiness, and on balance, most people are glad to have had the opportunity of existing. Given this, what could be immoral about creating ancestor-simulations? Since the inhabitants of an ancestor-simulation would feel the same way about their lives as we do about ours, mightn't it be immoral *not* to create ancestor-simulations, if one had the means of so doing?

But the situation is by no means this straightforward. The fact that simulations need not be unpleasant does not mean their creation is morally unproblematic:

> *The Objection from Lesser Value* A real life has greater intrinsic value than a subjectively similar simulated life. Since it is wrong to impose on others a low-grade form of existence that one would prefer to avoid oneself, the creation of simulated lives is immoral.

This objection may seem weak: even if virtual lives do possess less intrinsic value than their non-virtual counterparts, other things being equal, they can still be lives worth living, and hence lives that are worth creating. However, there is a further point to bear in mind. As Nozick's imaginary case of the experience machine reveals, the desirability of a life is not determined solely by the desirability of the experiences it contains. An experience machine will supply you with a (virtual) life of any kind you like, so by connecting yourself up to such a device you are guaranteed a very enjoyable (virtual) life, a life in which as many of your desires as you choose to come true will come true. But as Nozick observes, few of us would choose permanently to connect ourselves to an experience machine if we had the opportunity of so doing, and for good reason: 'What is most disturbing about them is their living of our lives for us' (Nozick, 1980, p. 44).

While this lesson is important, in the present context it is also of limited relevance. The virtual lives sustained by experience machines are of the *passive* kind: they consist of solitary streams of consciousness that are completely controlled and pre-programmed. As we have seen, not all virtual lives need be like this. Of particular interest here are *Diaspora*-style AC-simulations, i.e., virtual lives that are both *active* and *communal*, in the senses introduced above. Subjects in AC-simulations possess their own autonomous psychological systems (whether original or replacement). They lead their own lives: their actions are not pre-programmed (they are as free as anyone can be). And they can causally interact with other subjects in their virtual environment (and these other subjects are autonomous individuals in their own right, rather than merely the appearances of such). Given all this, it is hard to see why life in an AC-simulation should be regarded as being inherently less valuable or worthwhile than a normal life. True, the inhabitants of AC-simulations are not physically embodied in the normal way, but they can possess virtual bodies that are indistinguishable from the real thing. They are unable to manipulate physical objects, but they can manipulate virtual objects which *seem* physical. Why should the undetectable absence of a (non-virtual) material environment significantly diminish the value of the lives of these subjects? I cannot see any reason why it should.[29]

These considerations further weaken the Objection from Lesser Value. Even those who find this objection persuasive would only have reason to avoid creating passive simulations; there is no reason why they should be reluctant to create AC-simulations.

However, there is a further, and potentially more serious objection to the fostering of virtual life:

> *The Deception Objection* The subjects of simulations are being deliberately deceived; their lives are virtual, but they believe them to be real. This deception is engineered and maintained by the relevant simulators. Such actions are clearly wrong.

Deception is not an inevitable consequence of simulation; there may well be simulants who are perfectly aware of their true condition — as the software citizens in *Diaspora* are. But since few contemporary humans believe themselves to be leading simulated lives, the Deception Objection does apply to simulations of the menacing variety. This is not to say that it will have an impact on simulation policy. It is

[29] Berkeley was perhaps the first to make this point, when he argued for the redundancy of mind-independent material reality. The inhabitants of some of the polises in Egan's *Diaspora* take the same view.

conceivable that future simulators will take the view that although deception is wrong, the kind of deception being perpetrated on simulants does not constitute a wrong that is sufficiently serious to outweigh the boon of existence. But equally, it is conceivable that future simulators *will* be swayed by the Deception Objection, and restrict their simulation activities accordingly. This may not seem likely, but since we can only guess at the ways ethical considerations will influence the simulation policies of our (quite possibly super-intelligent) descendants it cannot be ruled out.

It should also be noted that the force of the Deception Objection may well depend on the type of simulation under consideration. The objection has considerable force in the context of long-term S-simulations of entire civilizations: anyone who creates an ancestor-simulation is responsible for the deceiving of billions of (virtual) people for thousands of (subjective) years. The situation is very different in the case of small-scale, short-term N-simulations. You are, let us suppose, feeling run-down by the demands of your 22nd century job, and decide to spend a couple of days in the (virtual) past to unwind; you employ the method of self-induced controlled hallucination, and 'wake up' in early 19th century England, in the midst of the Napoleonic wars. As you enjoy your adventure, are you the victim of a deliberate deception? In a sense, yes: you have opted for the fully-immersive trip, and so believe yourself to be a typical early 19th century person. But is the kind of deception involved in this case morally problematic? Surely not. Rather than one person imposing an uninvited delusion on another — as in the case of ancestor-simulations — we are dealing here with a person freely choosing to impose a short-term and harmless delusion *upon themselves*. Where is the wrong in that?

This implications of this point are by no means trivial, for as we saw earlier, given sufficient time, N-simulations might easily be created in sufficient numbers so as to be seriously menacing, even without the advent of superintelligence. But we are not yet done. There is at least one further reason why our descendants might avoid indulging in menacing simulatory practices, a reason that is pragmatic rather than ethical:

> *The Self-Interest Consideration* Future generations of humans and machines will be well acquainted with the simulation reasoning, and so will impose tight restrictions on simulation creation. They will realize that unless such restrictions are imposed, and enforced, no one — themselves included — will be in a position to rule out the likelihood that their lives are virtual rather than real.

I am not confident that such a policy will ever be adopted, for a number of reasons.

(1) Future simulators may well include superintelligent AIs. It is difficult to predict how these AIs will react to the knowledge that their own mental lives may well be simulations, nor can we predict — for the reasons outlined above—what their attitude to the creation of virtual human lives will be. Furthermore, even if we (humans) wanted to restrict the simulatory practices of these superintelligences, it is by no means clear that we would be able to.

(2) Leaving superintelligences aside, at a more mundane level, simulation technology is certain to play an increasing role in recreational activities, and people will become accustomed to, and demand, ever more lifelike simulations — just as today there is a demand for ever more life-like computer games. Since a ban on life-like simulations would be unpopular with both the public and powerful commercial concerns, the prospects of one being implemented are slim.[30]

(3) Many people will be unlikely to take the simulation argument seriously until they themselves have experienced what the technology can do, and taken a fully-immersive trip to the past or future. Should this point every be reached, billions of menacing simulations will have been created, and it will be obvious to everyone that it is already too late to consider a ban.

(4) To have the desired effect, a ban on simulations would have to be continued into the indefinite future. But even if an effective ban could be enforced in the present, we could never be confident that this policy would not be abandoned, or fail, at some future date — not least because it would be foolhardy to rule out the possibility of a singularity occurring, bringing with it hard-to-predict superintelligent machines. For this reason alone it is unlikely that our descendants would be willing to deprive themselves of all the benefits advanced simulation technology makes available.

There is a more general point. We are in the process of emerging from an age of innocence, an innocence that we are unlikely ever to recapture. Innocence was being able to believe that only sceptical possibilities of the most radical sort stood between ourselves and the world about us. This innocence evaporates on contact with the

[30] Perhaps I was not the only one to feel a slight chill when reading of Chalmers' discussions of the possible dangers of research into AI with West Point Military Academy cadets and staff. In short, there is no chance that the risks associated with an intelligence explosion are going to deter the military from trying to develop superintelligence: the dangers if the other side should get there first are simply too great (Chalmers, 2010, p. 29).

knowledge that even if reality is much as it seems, there is a significant likelihood that one's current consciousness is simulated. Having to live with this knowledge may well be part of the normal lot of technologically advanced conscious beings, whether biological or non-biological, the universe over. When this realization fully dawns on our descendants, attempting to recapture their lost innocence by imposing restrictions on simulatory practices will very likely strike them as futile. Since any restrictions on simulation creation can always be lifted subsequently, it will be obvious that their imposition would offer only meagre protection against the menace of simulation. But another factor will enter into the reckoning. Even if innocence once lost is impossible to regain, innocence can of course be *simulated*. If our descendants want to escape the shadow of simulation and experience for themselves what it was like to exist in more innocent times, they may have but one option: to embark on fully immersive virtual reality trips into the past. Not only does this further reduce the chances of restrictions on simulation creation being imposed, it is also bad news for our predecessors. It could easily be that the vast majority of people who find themselves living in more innocent times are simulants.

Our own predicament is only slightly better. Many of our descendants might be tempted by the prospect of finding out what it was like to *become* aware of the simulation menace; experiencing the first falling of the shadow might be an irresistibly appealing prospect. If so, life in the early 21st century may be an even more fragile thing than it appears.[31]

References

Banks, I.M. (2010) *Surface Detail*, London: Orbit.

Bostrom, N. (2003) Are you living in a computer simulation?, *Philosophical Quarterly*, **53** (211), pp. 243–255.

Bostrom, N. (2009a) The simulation argument: Some explanations, *Analysis*, **69** (3), pp. 458–461.

Bostrom, N. (2009b) *The Simulation Argument FAQ*, [Online], http://www.simulation-argument.com/faq.html

Burnyeat, M.F. (1982) Idealism and Greek philosophy: What Descartes saw and Berkeley missed, *The Philosophical Review*, **XCI** (1), pp. 3–40.

Chalmers, D.J. (1996) *The Conscious Mind*, Oxford: Oxford University Press.

[31] And you may feel it unwise to dwell on these matters further. My thanks to: Tim Bayne, Nick Bostrom, David Carlyon, David Chalmers, Stephen Clark, Richard Gaskin, Gerard Hurley, Jonathan Lowe, and audiences at Bradford, Glasgow and Stirling. I am also grateful to Durham's Institute of Advanced Study, which provided an ideal environment for thinkng about these issues.

Chalmers, D.J. (2005) The Matrix as metaphysics, in Grau, C. (ed.) *Philosophers Explore the Matrix*, Oxford: Oxford University Press. Also, [Online], http://consc.net/papers/matrix.html

Chalmers, D.J. (2010) The singularity: A philosophical analysis, *Journal of Consciousness Studies*, **17** (9–10), pp. 7–65.

Dainton, B. (2004) The self and the phenomenal, *Ratio*, **17** (4), pp. 365–389.

Dainton, B. (2008) *The Phenomenal Self*, Oxford: Oxford University Press.

Dainton, B. (2010) Temporal consciousness, *Stanford Encyclopedia of Philosophy*, [Online], http://plato.stanford.edu/entries/consciousness-temporal/

Dainton, B. (forthcoming) Selfhood and the flow of experience, *Grazer Philsophische Studien*.

Dainton, B. & Bayne, T. (2005) Consciousness as a guide to personal persistence, *Australasian Journal of Philosophy*, **85** (4), pp. 549–571.

Egan, G. (1997) *Diaspora*, London: Gollanz.

Flanagan, O. (2000) *Dreaming Souls: Sleep, Dreams, and the Evolution of Conscious Life*, Oxford: Oxford University Press.

Hanson, R. (2001) How to live in a simulation, *Journal of Evolution and Technology*, **7**, [Online], http://www.jetpress.org/volume7/simulation.htm

Johnston, M. (2010) *Surviving Death*, Princteton, NJ: Princeton University Press.

Lockwood, M. (1989) *Mind, Brain and the Quantum*, Oxford: Blackwell.

McGinn, C. (1999) *The Mysterious Flame*, New York: Basic Books.

Markham, H. (2006) The Blue Brain Project, *Nature Reviews Neuroscience*, **7**, pp. 153–160. Also, [Online], http://www.hss.caltech.edu/ steve/markham.pdf

Naselaris, T., Prenger, J., Kendrick, K., Oliver, M. & Gallant, J. (2009) Bayesian reconstruction of natural images from human brain activity, *Neuron*, **63** (6), pp. 902–915.

Nozick, R. (1980) *Anarchy, State, and topia*, Oxford: Blackwell.

Searle, J. (1992) *The Rediscovery of the Mind*, Cambridge, MA: MIT Press.

Strawson, G. (1994) *Mental Reality*, Cambridge, MA: MIT Press.

Susskind, L. (1994) The world as a hologram, [Online], http://arxiv.org/abs/hep-th/9409089

Wolfram, S. (2002) *A New ind of Science*, Champaign, IL: Wolfram Media Inc.

Zuse, K. (1970) *Calculating Space*, MIT Technical translation of *Rechnender Raum* (1969).