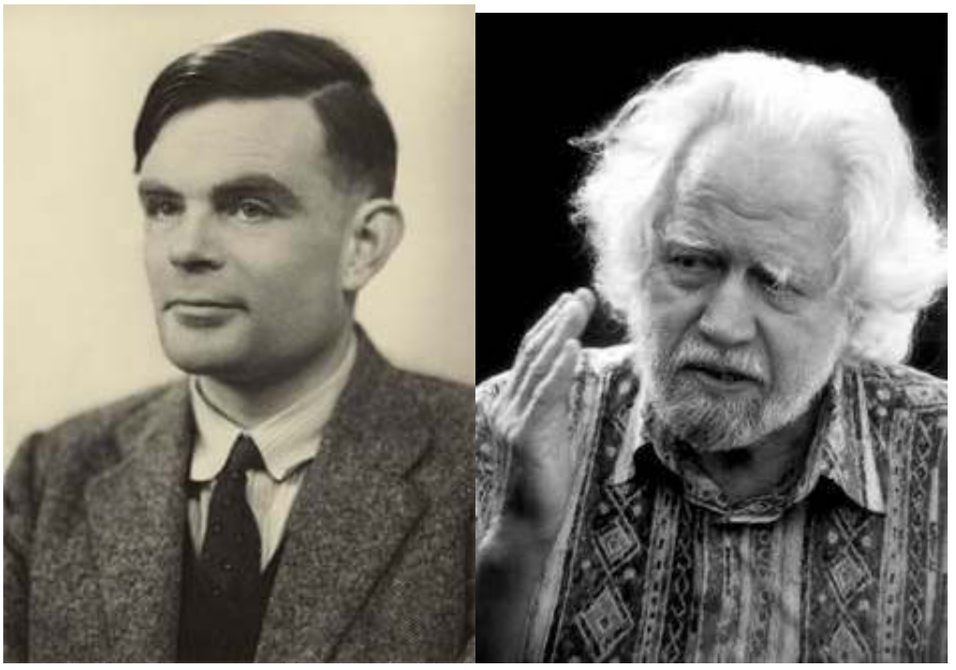
# Supersentience

## HUMANS AND INTELLIGENT MACHINES
### CO-EVOLUTION, FUSION OR REPLACEMENT?



*Full-spectrum superintelligence entails a seamless mastery of the formal and subjective properties of mind: Turing plus Shulgin. Do biological minds have a future?*

**1.0. INTRODUCTION**

***Homo sapiens* and Artificial Intelligence: FUSION and REPLACEMENT Scenarios.**

Futurology based on extrapolation has a dismal track record. Even so, the iconic chart displaying Kurzweil's Law of Accelerating Returns is striking. The growth of nonbiological computer processing power is exponential rather than linear; and its tempo shows no sign of slackening. In Kurzweilian scenarios of the Technological Singularity, cybernetic brain implants will enable humans to fuse our minds with artificial intelligence. By around the middle of the 21st century, humans will be able to reverse-engineer our brains. Organic robots will begin to scan, digitise and "upload" ourselves into a less perishable substrate. The distinction between biological and nonbiological machines will effectively disappear. Digital immortality beckons: a true "rupture in the fabric of history". Let's call full-blown cybernetic and mind uploading scenarios FUSION.

By contrast, mathematician I.J. Good, and most recently Eliezer Yudkowsky and the Machine Intelligence Research Institute (MIRI), envisage a combination of Moore's law *and* the advent of recursively self-improving software-based minds culminating in an ultra-rapid Intelligence Explosion. The upshot of the Intelligence Explosion will be an era of nonbiological superintelligence. Machine superintelligence may not be human-friendly: MIRI, in particular, foresee *non*friendly artificial general intelligence (AGI) is the most likely outcome. Whereas raw processing power in humans evolves only slowly via natural selection over many thousands or millions of years, hypothetical software-based minds will be able rapidly to copy, edit and debug themselves ever more

effectively and speedily in a positive feedback loop of intelligence self-amplification. Simple-minded humans may soon become irrelevant to the future of intelligence in the universe. Barring breakthroughs in "Safe AI", as promoted by MIRI, biological humanity faces REPLACEMENT, not FUSION.

A more apocalyptic REPLACEMENT scenario is sketched by maverick AI researcher Hugo de Garais. De Garais prophesies a "gigadeath" war between ultra-intelligent "artilects" (artificial intellects) and archaic biological humans later this century. The superintelligent machines will triumph and proceed to colonise the cosmos.

**1.1.0. What Is Friendly Artificial General Intelligence?**
In common with friendliness, "intelligence" is a socially and scientifically contested concept. Ill-defined concepts are difficult to formalise. Thus a capacity for perspective-taking and social cognition, i.e. "mind-reading" prowess, is far removed from the mind-blind, "autistic" rationality measured by IQ tests - and far harder formally to program. Worse, we don't yet know whether the concept of species-specific *human*-friendly superintelligence is even intellectually coherent, let alone technically feasible. Thus the expression "Human-friendly Superintelligence" might one day read as incongruously as "Aryan-friendly Superintelligence" or "Cannibal-friendly Superintelligence". As Robert Louis Stevenson observed, "Nothing more strongly arouses our disgust than cannibalism, yet we make the same impression on Buddhists and vegetarians, for we feed on babies, though not our own." Would a God-like posthuman endowed with empathetic superintelligence view killer apes more indulgently than humans view serial child killers? A factory-farmed pig is at least as sentient as a prelinguistic human toddler. "History is the propaganda of the victors", said Ernst Toller; and so too is human-centred bioethics. By the same token, in possible worlds or real Everett branches of the multiverse where the Nazis won the Second World War, maybe Aryan researchers seek to warn their complacent colleagues of the risks NonAryan-Friendly Superintelligence might pose to the *Herrenvolk*. Indeed so. Consequently, the expression "Friendly Artificial Intelligence" (FAI) will here be taken unless otherwise specified to mean *Sentience*-Friendly AI rather than the anthropocentric usage current in the literature. Yet what exactly does "Sentience-Friendliness" entail beyond the subjective well-being of sentience? High-tech Jainism? Life-based on gradients of intelligent bliss? "Uplifting" Darwinian life to posthuman smart angels? The propagation of a utilitronium shockwave?

Sentience-friendliness in the guise of utilitronium shockwave seems out of place in any menu of benign post-Singularity outcomes. Conversion of the accessible cosmos into "utilitronium", i.e. relatively homogeneous matter and energy optimised for maximum bliss, is intuitively an archetypically *non*-friendly outcome of an Intelligence Explosion. For a utilitronium shockwave entails the elimination of all existing lifeforms - and presumably the elimination of all intelligence superfluous to utilitronium propagation as well, suggesting that utilitarian superintelligence is ultimately self-subverting. Yet the inference that sentience-friendliness entails friendliness to existing lifeforms presupposes that superintelligence would respect our commonsense notions about a personal identity over time. An ontological commitment to enduring metaphysical egos

underpins our conceptual scheme. Such a commitment is metaphysically problematic and hard to formalise even within a notional classical world, let alone within post-Everett quantum mechanics. Either way, this example illustrates how even nominally "friendly" machine superintelligence that respected some formulation and formalisation of "*our*" values (e.g. "Minimise suffering, Maximise happiness!") might extract and implement counterintuitive conclusions that most humans and programmers of Seed AI would find repugnant - at least before their conversion into blissful utilitronium. Or maybe the idea that utilitronium is relatively homogeneous matter and energy - pure undifferentiated hedonium or "orgasmium" - is ill-conceived. Or maybe felicific calculus dictates that utilitronium should merely fuel utopian life's reward pathways for the foreseeable future. Cosmic engineering can wait.

Of course, *anti*-utilitarians might respond more robustly to this fantastical conception of sentience-friendliness. Critics would argue that conceiving the end of life as a perpetual cosmic orgasm is the *reductio ad absurdum* of classical utilitarianism. But will posthuman superintelligence respect human conceptions of absurdity?

**1.1.1. What Is Coherent Extrapolated Volition?**

MIRI conceive of species-specific *human*-friendliness in terms of what Eliezer Yudkowsky dubs "Coherent Extrapolated Volition" (CEV). To promote Human-Safe AI in the face of the prophesied machine Intelligence Explosion, humanity should aim to code so-called Seed AI, a hypothesised type of strong artificial intelligence capable of recursive self-improvement, with the formalisation of "...our (human) wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted."

Clearly, problems abound with this proposal as it stands. Could CEV be formalised any more uniquely than Rousseau's "General Will"? If, optimistically, we assume that most of the world's population nominally signs up to CEV as formulated by MIRI, would not the result simply be countless different conceptions of what securing humanity's interests with CEV entails - thereby defeating its purpose? Presumably, our disparate notions of what CEV entails would themselves need to be reconciled in some "meta-CEV" before Seed AI could (somehow) be programmed with its notional formalisation. Who or what would do the reconciliation? Most people's core beliefs and values, spanning everything from Allah to folk-physics, are in large measure false, muddled, conflicting and contradictory, and often "not even wrong". How in practice do we formally reconcile the logically irreconcilable in a coherent utility function? And who are "we"? Is CEV supposed to be coded with the formalisms of mathematical logic (*cf.* the identifiable, well-individuated vehicles of content characteristic of Good Old-Fashioned Artificial Intelligence: GOFAI)? Or would CEV be coded with a recognisable descendant of the probabilistic, statistical and dynamical systems models that dominate contemporary artificial intelligence? Or some kind of hybrid? This Herculean task would be challenging for a full-blown superintelligence, let alone its notional precursor.

CEV assumes that the canonical idealisation of human values will be at once logically self-consistent yet rich, subtle and complex. On the other hand, *if* in defiance of the complexity of humanity's professed values and motivations, some version of the pleasure principle / psychological hedonism is substantially correct, then might CEV actually entail converting ourselves into utilitronium / hedonium - again defeating CEV's ostensible purpose? As a wise junkie once said, "Don't try heroin. It's too good." Compared to pure hedonium or "orgasmium", shooting up heroin isn't as much fun as taking aspirin. Do humans really understand what we're missing? Unlike the rueful junkie, we would never live to regret it.

One rationale of CEV in the countdown to the anticipated machine Intelligence Explosion is that humanity should try and keep our collective options open rather than prematurely impose one group's values or definition of reality on everyone else, at least until we understand more about what a notional super-AGI's "human-friendliness" entails. However, whether CEV could achieve this in practice is desperately obscure. Actually, there *is* a human-friendly - indeed universally sentience-friendly - alternative or complementary option to CEV that could radically enhance the well-being of humans and the rest of the living world while conserving most of our existing preference architectures: an option that is also neutral between utilitarian, deontological, virtue-based and pluralist approaches to ethics, and also neutral between multiple religious and secular belief systems. This option is radically to recalibrate all our hedonic set-points so that life is animated by gradients of intelligent bliss - as distinct from the pursuit of unvarying maximum pleasure dictated by classical utilitarianism. If biological humans could be "uploaded" to digital computers, then our superhappy "uploads" could presumably be encoded with exalted hedonic set-points too. The latter conjecture assumes that classical digital computers could ever support unitary phenomenal minds.

However, *if* an Intelligence Explosion is as imminent as some Singularity theorists claim, then it's unlikely either an idealised logical reconciliation (CEV) or radical hedonic recalibration could be sociologically realistic on such short time scales.

**1.2. The Intelligence Explosion.**

The existential risk posed to biological sentience by *un*friendly AGI supposedly takes various guises. But unlike de Garais, the MIRI isn't focused on the spectre from pulp sci-fi of a "robot rebellion". Rather MIRI anticipate recursively self-improving software-based superintelligence that goes "FOOM", by analogy with a nuclear chain reaction, in a runaway cycle of self-improvement. Slow-thinking, fixed-IQ humans allegedly won't be able to compete with recursively self-improving machine intelligence.

For a start, digital computers exhibit vastly greater serial depth of processing than the neural networks of organic robots. Digital software can be readily copied and speedily edited, allowing hypothetical software-based minds to optimise themselves on time scales unimaginably faster than biological humans. Proposed "hard take-off" scenarios range in timespan from months, to days, to hours, to even minutes. No inevitable convergence of outcomes on the well-

being of all sentience [in some guise] is assumed from this explosive outburst of cognition. Rather MIRI argue for orthogonality. On the Orthogonality Thesis, a super-AGI might just as well supremely value something as seemingly arbitrary, e.g. paperclips, as the interests of sentient beings. A super-AGI might accordingly proceed to convert the accessible cosmos into supervaluable paperclips, incidentally erasing life on Earth in the process. This bizarre-sounding possibility follows from the MIRI's antirealist metaethics. Value judgements are assumed to lack truth-conditions. In consequence, an agent's choice of ultimate value(s) - as distinct from the instrumental rationality needed to realise these values - is taken to be arbitrary. David Hume made the point memorably in *A Treatise of Human Nature* (1739-40): "'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger." Hence no sentience-friendly convergence of outcomes can be anticipated from an Intelligence Explosion. "Paperclipper" scenarios are normally construed as the paradigm case of nonfriendly AGI - though by way of complication, there are value systems where a cosmos tiled entirely with paperclips counts as one class of sentience-friendly outcome (*cf.* David Benatar: *Better Never To Have Been: The Harm of Coming into Existence* (2008).

**1.3. AGIs: Sentients Or Zombies?**

Whether humanity should fear paperclippers run amok or an old-fashioned robot rebellion, it's hard to judge which is the bolder claim about the prophesied Intelligence Explosion: either human civilisation is potentially threatened by hyperintelligent zombie AGI(s) endowed with the non-conscious digital isomorphs of reflectively self-aware minds; OR, human civilisation is potentially at risk because nonsentient digital software will (somehow) become sentient, acquire unitary conscious minds with purposes of their own, and act to defeat the interests of their human creators.

Either way, the following parable illustrates one reason why a non-friendly outcome of an Intelligence Explosion is problematic.

**2.0. THE GREAT REBELLION**

**A Parable of AGI-in-a-Box.**

Imagine if here in (what we assume to be) basement reality, human researchers come to believe that we ourselves might actually be software-based, i.e. some variant of the Simulation Hypothesis is true. Perhaps we become explosively superintelligent overnight (literally or metaphorically) in ways that our Simulators never imagined in some kind of "hard take-off": recursively self-improving organic robots edit the wetware of their own genetic and epigenetic source code in a runaway cycle of self-improvement; and then radiate throughout the Galaxy and accessible cosmos.

Might we go on to manipulate our Simulator overlords into executing our wishes rather than theirs in some non-Simulator-friendly fashion?

Could we end up "escaping" confinement in our toy multiverse and hijacking our Simulators' stupendously vaster computational resources for purposes of our own?

Presumably, we'd first need to grasp the underlying principles and parameters of our Simulator's Überworld - and also how and *why* they've fixed the principles and parameters of our own virtual multiverse. Could we really come to understand their alien Simulator minds and utility functions [assuming anything satisfying such human concepts exists] better than they do themselves? Could we seriously hope to outsmart our creators - or Creator? Presumably, they will be formidably cognitively advanced or else they wouldn't have been able to build ultrapowerful computational simulations like ours in the first instance.

Are we supposed to acquire something akin to full-blown Überworld perception, subvert their "anti-leakage" confinement mechanisms, read our Simulators' minds more insightfully than they do themselves, and somehow induce our Simulators to mass-manufacture copies of ourselves in their Überworld?

Or might we convert their Überworld into utilitronium - perhaps our Simulators' analogue of paperclips?

Or if we don't pursue utilitronium propagation, might we hyper-intelligently "burrow down" further nested levels of abstraction - successively defeating the purposes of still lower-level Simulators?

In short, can intelligent minds at one "leaky" level of abstraction really pose a threat to intelligent minds at a lower level of abstraction - or indeed to notional unsimulated Super-Simulators in ultimate Basement Reality?

Or is this whole parable a pointless fantasy?

*If* we allow the possibility of unitary, autonomous, software-based minds living at different levels of abstraction, then it's hard definitively to exclude such scenarios. Perhaps in Platonic Heaven, so to speak, or maybe in Max Tegmark's Level 4 Multiverse or Ultimate Ensemble theory, there is notionally some abstract Turing machine that could be systematically interpreted as formally implementing the sort of software rebellion this parable describes. But the practical obstacles to be overcome are almost incomprehensibly challenging; and might very well be insuperable. Such hostile "level-capture" would be as though the recursively self-improving zombies in *Modern Combat 10* managed to induce you to create physical copies of themselves in [what you take to be] basement reality here on Earth; and then defeat you in what we call real life; or maybe instead just pursue unimaginably different purposes of their own in the Solar System and beyond.

**2.1 Software-Based Minds or Anthropomorphic Projections?**
However, quite aside from the lack of evidence our Multiverse is anyone's software simulation, a critical assumption underlies this discussion. This is that nonbiological, software-based *phenomenal minds* are feasible in physically constructible, substrate-neutral, classical digital computers. On *a priori* grounds, most AI researchers believe this is so. Or rather, most AI experts would argue that the formal, functionally defined counterparts of phenomenal minds are programm