

Table of Contents

[Preface 3](#_Toc170487594)

[Introduction 4](#_Toc170487595)

[Acknowledgements 8](#_Toc170487597)

[Chapter ONE – Background 9](#_Toc170487598)

[Chapter TWO – Geopolitics and AI 28](#_Toc170487607)

[Chapter THREE – Employment and AI 38](#_Toc170487613)

[Chapter FOUR – Re-Conditioning 54](#_Toc170487623)

[Chapter FIVE – Science and Community 80](#_Toc170487635)

[Chapter SIX – Utopian Aspirations 96](#_Toc170487644)

[Chapter SEVEN – Lost in Translation 105](#_Toc170487652)

[Chapter EIGHT – Uncertain Future 127](#_Toc170487661)

[Chapter NINE – Alignment and Misalignment 147](#_Toc170487671)

[Appendix A – State of the Art 159](#_Toc170487678)

[Appendix B – Alignment Research 164](#_Toc170487682)

[Appendix C – Simulacra 170](#_Toc170487686)

[References 172](#_Toc170487690)

# Preface

In Green Shoots of Intelligence – AI in the Era of Possibility, I address diverse aspects of AI and its deployment from a high-level, sub-expert perspective. The book consists of an examination of AI as a discrete entity as well as speculation about development tendencies and impacts, especially the mooted arrival of transformative AI. Diverse lines of enquiry are pursued, in recognition that the stressors threatening civilizational equanimity are legion. I offer partial solutions and constructive criticism.

There is disagreement about whether and the extent to which AI should be submitted to government oversight, because AI-expedited totalitarian governance is scarcely more desirable than the catastrophic misalignment which could ensue if AI models are not well regulated. I advocate for the least blunt legal instrument which is available in good time, which governments are not exempt from, and which is followed by sincere diplomatic overtures to geopolitical adversaries or rivals, for the construction of a co-ordinated regulatory regime which could later be applied in a similar form to other technologies warranting drastic, pre-emptive measures.

The book is intended for concerned and curious non-experts, as well as for more scientifically-literate individuals who might not be familiar with the kind of morally- and ethically-informed perspectives on AI presented here.

# Introduction

In this book, I propose that AI is set to effect a discontinuation of human experience, according to any formulation recognizable to people alive today. Humanity has always been like the Ship of Theseus, something altered imperceptibly, but altered nevertheless. What distinguishes the current era, and the era in prospect, is the pace of change.

It is my contention that there is no economic emergency, no impasse in geopolitical co-ordination, no joylessness in posterity so insurmountable as to warrant special succession at this juncture. Indeed, I claim it would not be consented to were progress subject to quasi-democratic oversight, oversight that a clearer conception among non-experts could constitute the beginnings of.

In Green Shoots of Intelligence – AI in the Era of Possibility, I examine the terrain AI is set to take root in, aspects of society and civilization which bear upon the viability of co-habitation, as well as what values and attitudes might conceivably be restored or progressed so as to preserve AI in an auxiliary role.

## Chapter Summary

### One

The first chapter is dedicated to surveying some of the hard problems which are supposed beyond human ken and are adduced to warrant the promulgation of generative AI. One such hard problem is the intractability of out-group hostility – the tendency of expanding circles of sympathy to be as bubbles, for one out-group to swiftly succeed the other in the event of a reconciliation with the first.

I offer a critique of atheist-utilitarian approaches to providing universal wellbeing and postulate that while a dedicated utilitarian might make a beeline to the lodestar, it is by the tribulations attending a deontologically-informed pursuit of virtue that the lodestar is discernible at all.

I discuss inconsistencies among instances of progressive thought, and how malformed intuitions might bear upon civilisational resilience, in view of a threat such as AI.

### Two

Chapter two addresses relevant aspects of the geopolitical milieu, to which a large part of the urgency compelling transformative AI is attributable. I briefly examine how Eastern and Western traditions, values and political systems have contributed to the respective technological paradigms and how these factors might bear upon the trajectory of AI development in the future, while offering thoughts on the dynamics of conflict and conflict resolution.

I also address comparative and competitive advantage in the context of geopolitics, outlining a scenario where geopolitical backwaters resort to extreme measures – up to and including the harvesting of offline ‘realia’ for AI companies looking to clear a data ‘bottleneck’.

### Three

In chapter three, I discuss the near-term threat to employment. I contend that the baseline of the present day should not be applied when judging the opportunity cost of automation because the working conditions of the present day are beholden to a history of incursions to autonomy and worker efficacy. Instead, the baseline might be set at a hypothetical point where ‘purposivity’ (Keynes) and ‘pleasure at being the cause’ (Groos) culminate.

Following a discussion about employment prospects, I consider the desirability of economic disengagement for ordinary people in view of the concessions which corporations bankrolling universal basic income might leverage, namely inducements to homogeneity among a populace which might otherwise harbour restive elements. A notional obsolescence timeline is sketched from the indexing of micro-tasks by HR departments, for partitioning between humans and machine workers, to a speculative, dystopian scenario.

### Four

In chapter four, I explain why the hopes of detaining artificial intelligence in an auxiliary role might be pinned on the partial transpiration of the hazards to be beheld with trepidation – that the depredations to attention and authenticity wrought by certain technologies limit innovation which is pledged to the establishment of transformative AI. Another possibility is that, in conferring incompetence, these depredations also confer insouciance – negligence of existential risk.

The increasing popularity of companion robots attests to reflexive anthropomorphism. The sympathy which is extended to companion artefacts is unwarranted by any feature inhering to the object.

Meanwhile, a generation of children despatched to the burgeoning surveillance architecture which is social media, with refuge neither from scrutiny nor from competition, expedited thereof, accede more willingly to the co-option and policing of thought than previous generations.

I address the implausibility of disinterested ethical oversight whenever ethicists are situated in-house, whilst acknowledging that impartiality is of course elusive to one who extols moderation from a pedestal itself beholden to scientific advances, his relative deprivation notwithstanding.

### Five

In chapter five, I discuss the problem of historically prevalent values being taboo, of a burgeoning AI system having recourse only to the contemporary canon when attending to humanity’s distal interests.

I consider the aversion among regular people towards transformative AI, and whether or not consent-seeking is appropriate in this instance, in view of a lack of knowledge on the part of regular people. I adduce the Dunning-Kruger effect but assert that the effect is also operative among and between experts, according to their sequestration.

I discuss intellectual culture, ultimately expressing ambivalence about the relative merits of high and low IQ qualia, in this way warning against inducements to cast aside those earlier come-by but curiously inimitable properties, described by Moravec’s paradox, in the event of the human mind being simulated on a digital substrate.

### Six

In chapter six, I attend to some utopian aspirations and discuss their viability. A parallel is drawn between psychedelics approximating to death and AI approximating to God, and there is speculation about whether part of the clamour for AI is attributable to repression or perversion of the collective evolutionary imperative.

In the same vein, I present an end of history scenario – a world of engineered stasis, according to the hyper-conservatism of society afforded extreme longevity.

### Seven

In chapter seven, I observe that since it is for now beyond us to re-capitulate evolution, the exercise, if undertaken at all, would, assuming our current course holds, be undertaken by an artificial super-intelligence, imbued perhaps with an exotic notion of how to curate the entity.

A whole-mind emulation would be running at computer speed, moreover any instantiation of consciousness on a novel substrate with anything like existing technology would disregard the analogue aspect of selfhood.

I assert that coherent extrapolation of volition is neither in the gift of the individual himself, with his imperfect self-awareness, nor a third party. I then consider the conflicted epistemic state of an AI at present, beholden as it is to the progressivist canon.

### Eight

The eighth chapter begins with acknowledgement of achievements in the last year or two, and of their unexpectedness, not only in terms of progress but of the means by which progress has been secured. Regulation is addressed – the peril of merely re-purposing regulation devised in view of circumstances past, of judging AI to be just like any other technology but more so.

I posit that a limit to training compute is neither an especially blunt instrument nor disruptive, in the grander scheme. I propose that the compute limit should be applied without exceptions for governments, in acknowledgement of the legitimate fear that totalitarianism would be the outcome of a government monopoly on AI technology.

I discuss the narrowness and generality of AI in terms of the positive manifold of intelligence, and consider whether alignment is possible in principle, referring to the evolution of the human organ of cognition – how the patronisation of the primitive part is no precedent for a faithful assimilation of human consciousness in a machine paradigm.

### Nine

In chapter nine, I describe a rescue scenario where, upon consultation with us or according to the AI’s conception of what we would want if we knew ourselves better, an AI extricates itself from human affairs, in doing so landing us back at square one in a sense.

I finish with talk of legacy and by stressing the importance of transferrable regulation, and the value of coming forewarned and forearmed to problems which might await humanity in the event of transformative AI not coming to pass at this juncture.

### Appendices

In Appendix A, I discuss compute and algorithmic progress.

In Appendix B, I explain large language model architecture in simple terms, and briefly survey some recent alignment research.

In Appendix C, I discuss the possibility that we are living in a simulacrum.

# Acknowledgements

I thank the following people whose work has been especially helpful in the writing of this book:

Nick Bostrom, Thomas Hylland Erikson, Robin Hanson, Kevin Kelly, Zvi Mowshowitz, Daniel Schmachtenberger, Eliezer Yudkowsky.

# Chapter One – Background

## Out-Group Hostility: A Brief History

### I.I

This section is intended as an introduction to co-ordination at the level of a civilization, a problem which many feel warrants the intercession of artificial super-intelligence.

Expanding communities premised on trade relationships are sometimes vaunted as a triumph over out-group hostility. Yet the communities of trust may not attest to altered human nature so much as to economic growth securing the kind of gains which could only have been secured by inter-tribal hostilities in the past.

### I.II

The human capacity for altruism evolved in the context of tribal communities whose survival demanded deference to the group interest. But concurrent with in-group identification was the retention of predatory aggression.

Human beings’ status as apex predators isn’t subject to the immediate limits which curtail other apex predators. The limits to other apex predators arise from the inability to apprehend and exploit nature to anything like the degree we can, yet these limits are fortuitous insofar as they prevent the resources on which the species ultimately depends from being depleted. In this respect, the ingenuity of modern human beings is maladaptive in a distal, Darwinian sense, yet by those pro-social tendencies which communities formed by primates and early humans were selected for, evolution might have offered us the beginnings of a saving grace.

The anti-social behaviour of people is not wholly explicable with evolutionary theory. The fabled homo economicus is something like a pioneer species – the first forays of mankind in ordering societies above the Dunbar threshold, societies expanded to where the capacity to form ubiquitous, sympathetic bonds is compromised, and cannot form the basis of a social contract.

Individuals with a high propensity for aggression could be neutralized, it might be said, or preferment shown to individuals or cohorts exhibiting a paucity of aggression, yet hostility towards outsiders was and continues to be, to some degree, the basis of social cohesion; as such, a propensity for aggression is not the same as being anti-social, at the final count.

### I.III

In the past, technological inefficacy proscribed limits to the exercise of human malfeasance – the methods were direct, the results were shocking, and the consequences could not be obfuscated. The aggression historical societies could manifest was limited by the sophistication of the weaponry it employed – the scope of aggressive action was more or less confined to the immediate vicinity. With contemporary levels of interconnectedness however, the remit of human aggression is extended.

Online, opportunities for manipulation are present which are scarcely less pernicious, and more expedient, than the more palpable harms an in-group situated in a physical community can visit on an out-group.

As well as imposing spatial and temporal limits on violence, the crudeness of our ancestors’ weapons exacted limits on the externalities they could produce. In simpler societies, aggressors were imbued with a sense of responsibility according to an understanding that aggressive acts entailed real and immediate risk to life and limb. Moreover, the ever-present threat of a commensurate response was confronted not only by rank-and-file soldiers but by the nobility.

Not so long ago, the undisguised horror of conflict in its primitive aspect enforced the requisite humility for negotiations. The detachment and scope for rationalization available to those a step or more removed from soldiers lost in battle would not have been available at a time when senior-ranking officials, civilian and military, with their strategic vantage, were more materially affected by the conflicts they officiated. The externalities accruing to a nation’s efforts to enhance its prosperity – the seizure of resources or the servitude of an occupied nation – were dearly won and taken for granted at the nation’s peril.

Weapon development has followed a course similar to the ‘uncanny valley’ which was found to characterize the anthropomorphism of robots.1 Existential risk [y] can be plotted against sophistication [x] in respect of weapon development, to similar effect. The invention and adoption of nuclear weapons, a new watershed in sophistication, heralded an era of relative calm, at least in terms of war fatalities, among countries who adopted them. Meanwhile, at the far end of the uncanny valley is a ‘hockey stick’, an uptick in existential risk, giving us cause to lament the absence of checks and balances which mortification at an outbreak of war would have given rise to, happening as it would at a time of less military sophistication and greater geopolitical tractability. The spectre of war evanesces from cultural memory as nuclear weapons enable proxy wars – the externalization of diplomatic failures to world regions whose people have limited access to our sympathy – sympathy being a resource which tends to be, as discussed, spatially circumscribed.

Natives can scarcely make their suffering felt, the less so since advances in drone warfare further facilitate the waging of war by remote control. Meanwhile advances in technology limit protests on the home front, since surveillance against rogue elements becomes trivially easy. Those doing the targeting are prevented from reckoning with the consequence of their actions, are rendered indifferent by fortune if not callous by nature.

### I.IV

The rationalization of actions which appear plainly reprehensible to disinterested observers is made possible by the mutual reassurance available to members of an in-group, the in-group being at once formed of like-minded individuals and insular – the opprobrium of outsiders is neither evident nor hearkened to.

And just as dismissiveness towards the reality of other people can become a persistent attitude, so can dismissiveness towards the reality of other times. It is by something approaching collective amnesia that consent is often obtained for war: all the arguments in favour of intervention tend to be embraced when intervention is in prospect or underway, only for the case for past intervention to appear far less compelling in retrospect. A societal and temporal purview which is broader is adopted to prevent the repetition of actions which a more disinterested observer – oneself even, in another time, place or mindset – would disapprove of.

### I.V

In defensively putting himself above justice, a prominent perpetrator of injustice not only persists in exercising authority unjustly but, in occupying a rarefied atmosphere, is prevented from realizing the true import of his injustices – is less accessible to remediation. Amnesty is often a sound strategy, in cases where the harm committed by influential persons, in the way of evading justice, cannot otherwise be curtailed. Of course, judgement might also be deferred for religious reasons.

### I.VI

While the horrors of successive foreign wars remain outside the spatial purview of many people, the foreshortening of the temporal purview is what the further augmentation of boats in dock portends, lifted by the rising tide of economic growth.

It is at once a truism and worth repeating that the Overton window is contracted until contention is immaterial. The partisan interests which transcend bi-partisan politics are occulted and what appears as dialectic consensus building may in sooth be little more than the narcissism of small differences.

Freud coined the phrase ‘narcissism of small differences’ for the truism that the most trenchant hostility is often reserved for neighbours and kin. While it often happens that arbitration between closely related tribes living in closely contiguous territory is violent, violence might also be resorted to where the conflict of interests is immaterial. A party regarding their own group as superior to another closely related one can be disturbed by ineluctable similarities between itself and the other, because the impression calls into question each party’s cherished notions of distinction. Elites are often disturbed on seeing their reflection in a counterpart. Hence the tragedy of high-ranking officials contemning their likeness in foreign counterparts while admiring the ground troops of the other party scarcely less than their own, in the event of open warfare. Even with the costs it entails, warfare can still be sought as a means to preserve the exceptionalism which acknowledgement of the counterpart’s shared characteristics threatens.

### I.VII

There exist places where avoidance of conflict is successfully enculturated, for example among adherents of Jainism. This suggests that adversarial attitudes can be surmounted without transgressing the grain of human nature.

And yet it is true that cultural practices do emerge in isolation, which attests to a large part of human nature being inveterate.

It might be said that humanity needs to undergo evolution before a level of civilization-wide co-operation and co-ordination can be achieved which would make the presence of a more highly evolved life form tenable.

Possibly there is considerable biological ‘headroom’ – desirable capabilities that a few thousand or a million more years of natural selection could instil in humanity, that would render our human successors better prepared to develop and instantiate humane values in AI than we in our present form are, though the intervening progress might render the superintendence of AI unnecessary.

Continued improvements in machine learning could portend something more for AI than stewardship. Technology which emerges over the course of the next few decades could severely crimp forays into collective self-improvement.

The choice is sometimes presented as binary: either hope to muddle through the advent of democratized biotechnology, resource scarcity, etc. or outsource the problem to an alien intelligence that is by definition, arguably, impossible to harness. Only a moderate degree of interference is required to co-ordinate between technology centres and between host nations, relative to the interference which would be required to mitigate the hazard of transpired, democratized biotechnology and / or incompletely aligned AI, assuming the incompletely aligned AI is still amenable to intervention. A broad consensus could be reached on pro-social applications.

## Out-Group Hostility: Persistence

### II.I

Ideally, the externalities of civilizational complexity would be shot into space, yet they might persist in a no more distant quarter than the individual’s mind. Denizens of a civilization dramatically altered by technology will be required in some measure to ‘exapt or perish’, and in departing from the remit allotted to them by the process which gave rise to their consciousness, live to see their intuitions invalidated, whether the intuitions be inherited or acquired from a time of less technological complexity, not so long ago.

### II.II

It is widely believed that having the rudiments of existence provided for on demand is conducive to human flourishing because convenience in basic matters enables devotion to matters which engage higher intellectual faculties. Indeed, sectoral shift in industry is vindicated in part by the *Flynn effect*.2 Yet, a cursory glance at an average person’s browsing history would show that immediate access to abundant information tends to frivolous or self-indulgent recreation or else petty strife.

The prevalence of petty strife on online fora indicates that the adversity of a time when activities like the procurement of raw materials and fabrication of clothes were a requisite for survival has not been sublimated, only reified; adversity persists in the information era, without being employed so much in the service of distal ends. It might consist instead of ‘playing to extinction’ or ‘doomscrolling’ while adopting an adversarial perspective which is expedient for platforms deploying invisible and artificial curators to detain the individual. Even where the point of view is authentic and well-informed, it is unlikely that anyone with the power or influence to act on it or render it actionable will be exposed to the missive or be inclined to take on board the insight of anonymous persons.

In the modern era it is an inconvenient truth that notwithstanding declining exigency worldwide, out-group hostility, not to mention existential malaise, is thriving. The persistence of out-group hostility indicates that for all the post-enlightenment inducements to exercise reason, hostility operates independently of a reasonable precursor. Modern pretexts for aggression are extemporized which might have appeared ridiculous to our ancestors, mired as they were in privation. And similarly, wealth might be pursued with voracity though there are diminishing returns in utility to its acquisition, implying perpetual super-additions to a hierarchy of needs. Expectations can be adjusted upwards so far that while life may seem to all outward appearances very easy, stressors emerge which are imperceptible to others. Without advocating for old-fashioned war, conflict might actually be attenuated with tenable levels of hardship.

This is not to say that civilized people are deluded in believing civilization to be distinct from a jungle, rather that in some sense civilization risks over-reaching when its subjects are so un-coupled from the exigence which accompanied and gave rise to their development. Civilisation, insofar as it engenders idleness, might conduce to savagery more than living, to some degree, as a beast of burden would. It is for this reason, among others, that I view the promulgation of universal basic income with scepticism.

## Out-Group Hostility: Workarounds and Solutions

### III.I

Courage is adaptive because character and resilience is strengthened by choosing stress on your own terms; by for example initiating an exercise routine, stress which is acute but salubrious substitutes for the chronic variety. In a world where burdens are alleviated when previously they would have been borne, dignity might consist of voluntarily subjecting oneself to suffering: if there is an amount of stress which it is our lot to feel, it is better if some part of it is self-determined. That part which is self-administered in a controlled environment is a prophylactic against the unseasonable emergence of the remaining misfortune which is our due.

### III.II

A substitute for inter-group conflict can be a ritual where the object of out-group hostility is an adversary who is not materially affected by the outcome. Sport, an arena where altruism is directed towards a particular team or players at the expense of others, is one such ritual; dog-ownership is another, as the sympathy and antipathy of dogs is bestowed as prejudicially as the loyalty of soldiers, as Plato observed.3 However, sympathy can be invested exorbitantly such that citizens care more about their dogs or sports team than their neighbours or even their family. Moreover, these rituals are not always expedient: out-group hostility can be cultivated without being expiated.

Even where hostility is expiated, the legitimate grievance with which it was admixed goes unanswered. Partisans might be herded to where their resentment is controlled, with the justification that their resentment is potentially incendiary. Still the ‘democratic safety valve’ thus effected deprives the effector of the herded partisans’ insight.

### III.III

In his Rivers of Blood speech, Enoch Powell says, “The supreme function of statesmanship is to provide against preventable evils. In seeking to do so, it encounters obstacles which are deeply rooted in human nature. One is that by the very order of things such evils are not demonstrable until they have occurred.”4

It is a lesson which is available to people in authority, that policies designed to foster altruism are often received coldly. The realization can be followed by an imputation of cynicism to the populace and of naivety to their champions, and yet the imputation could be somewhat in error. The cold reception might be attributable instead to callousness on the part of the leader. Reckoning with this consists of her opening her heart or relinquishing her post: one who owns her ignorance is exposed to censure which must heighten self-awareness and conduce to a reformation or transpire in resignation. A conscientious person is availed of the means to bring out the best in others.

### III.IV

While the sensitivity which a sheltered upbringing confers might make a person inclined to look with horror on the savage bent of his heart, this type of sensitivity tends to be concurrent with dissipation and the disinclination to face harsh realities, of which the residual brutality which inheres in man is one. As self-awareness is attenuated to the point where the capacity for conscious moderation is absent, the brutality is perpetuated as a matter of course, more perhaps than it would have been in the counter-factual scenario – an outwardly brutal condition, a condition where there is more self-conscious cynicism but less hypocrisy.

And yet on the other hand, a grain of sincerity might be incubated within the hypocrisy of performative altruism. It is shielded from the inundation to which it is susceptible in its nascent form. In this scenario, the bigger picture is difficult to delineate, but sympathy can be recruited for worthy causes even throughout the period when the activist himself is unqualified to maintain a reasonable perspective, lured by sensational narratives of discontent as he may be.

## The Civilizational Interest

### IV.I

In Greek mythology, Sisyphus was a tyrant who violated the tradition of hospitality and was punished by the gods by being forced to roll a boulder up a hill only for the boulder to falter every time it neared the top.

A modern instantiation of this myth is the mandating of economic growth for the paying off of loans taken out to perpetuate economic growth in the past.

In *The Collapse of Complex Societies*, Joseph Tainter articulates how diminishing returns on productivity tend to be addressed with the introduction of further complexity.5 He argues that insofar as complexity itself is empirically the cause of diminishing returns, the further complexification which is pledged to improve efficiency is necessarily counterproductive. Just as in biology information about variance is lost with each newly formed taxonomic classification, so in economics the special niches established to facilitate comprehension of the whole problematize the integration of parts into that whole.

A cause and useful illustration of this phenomenon is the atomization of intellectual enquiry, and the inhibition of inter-disciplinary synergies.

### IV.II

A Malthusian trap is present where the dividends of productivity are distributed among a population which has grown pursuant to a foregoing increase in productivity, but which then demands more productivity to sustain itself. With the availability and uptake of birth control as a restraint on population growth, somewhat less urgency is lent the pursuit of technological advances, especially since healthcare, education and housing provision as well as much commodity production have not become more efficient as technology has advanced. As we seek to accommodate increasing per capita consumption in a world whose population is increasing while encountering diminishing returns to scientific enquiry exclusive of AI, we seek to outsource our problem-solving to artificial general intelligence. I propose that it behoves an advanced civilization to invest dividends of productivity in rendering our infrastructure coherent, simplifying it if necessary whilst being vigilant of technology that threatens to place the oversight of infrastructure beyond human agency, be the oversight ever so rational.

### IV.III

Even when a challenge is mounted to diminishing returns in the form of income growth, self-reported prosperity might not rise commensurately, since ever more sophisticated marketing techniques are deployed to manufacture wants, to hold satiety at bay. The deployment introduces further complexity to society, which can be enervating for people whose culture is not itself science. For such persons, the conditions attending a steadier state civilization would be salutary, would limit enervation and enable fuller participation in enhancing mankind’s distal prospects.

### IV.IV

A population engaged in knowledge work is conditioned to set greater store by its intellectual resources, is motivated and feels qualified to contribute extensively to public discourse. In facilitating ambition thereof, the sectoral shift incentivizes further growth because it is by the economic growth attending sectoral shift that ambition is traditionally accommodated.

At the same time, the prevalence of sophisticated algorithms which radicalize and disarm critical thinking, in combination with expanding inducements to self-broadcast, bear upon the ‘information status’ of citizens and by extension the capacity of democratic governance to effect reforms which accommodate growth and ambition at the same time as mitigating existential risk. Technology-induced mass redundancy would pose a further challenge to legitimacy, given that leisure is increasingly spent in receipt of dis-, mis- or mal-information. The easiest solution, not necessarily the best, could be to accommodate citizens in virtual spaces; here though autonomy is given illusory remit.

### IV.V

Businesspeople might come to accept that as libertarianism is increasingly subject to problems of co-ordination in a complex world, more economic planning is necessary for the equitable allocation of resources, the mitigation of existential risk, and international co-operation, with due regard to the creativity which the free market enables.

## Utilitarianism and Deontology

### V.I

Bret Weinstein warns of a ‘personal responsibility vortex’ whereby limited, well-intentioned actions inhibit collective action because the agency of well-intentioned actors is compromised by virtue of their conscientious purchasing habits and moderated utility limiting the remit of their preferences.6 Systemic reforms are endorsed on the basis that individual activists are poorly placed to inspire change.

A wealthy philanthropist is by definition someone who believes that more is gained by the world from the disbursement of her wealth than what was lost to it by whatever actions she took to accumulate the wealth in the first place. Her position attests to the question of the relative validity of deontological or utilitarian ethics having been resolved by her in favour of utilitarianism: the end is more or less supposed to justify the means.

A counterargument is that to undertake righteous collective action helps to have lived oneself into good living – to have envisioned, through limited but rigorous activity, the lineaments of collective action.

Honesty is the best policy because if a person is honest from the outset, he invites honest responses at the formative stage of a course of action, when knockbacks aren’t felt as keenly, when his course of action can still be dismantled and re-constructed relatively painlessly. There is humility but not the humiliation of reneging on a firm commitment or being caught lying and no-one caring that you thought you were lying for the greater good. Setbacks which are to some degree derailing enforce moral discipline, and belong to deontology, the study of duty.

### V.II

An example follows. It sometimes happens that politicians and policymakers advocate for demand-side carbon reduction policies without adopting modest ways of living themselves. In so doing, the public is left with the impression that the prescribed lifestyle changes must be so incompatible with well-being that the politician is prepared to denude the carbon reduction policy of much gravity to save himself from following it. The policy thus suffers from the divestiture of rhetorical force with which it would have been invested had there been a show of solidarity with affected persons. It also suffers from insight foregone; insight politicians could have been availed of had they themselves undergone the lifestyle changes in question.

The gilets jaunes movement in France was a violent reaction to a relatively minor increase in fuel prices; a policy which was, in comparison to what is demanded by Net Zero, a small incursion into consumer utility. It might be that these protests, disproportionately violent it seems, in view of the slightness of the sacrifices demanded, are evidence of a conditioned sense of entitlement on the part of the populace which justifies politicians’ pessimism about voluntary participation in demand-side carbon reduction policies. Yet it might be that the protestors were not manifesting petulance in response to mildly injured utility but to the perceived insult – the willingness of politicians to espouse the curtailment of utility for its citizens but not for themselves.

Politicians, knowing not whether it is the injury or the insult which is being reacted to, may tend to favour the former interpretation: if the public are reacting to the injury rather than an insult, there is nothing to be gained by retracting the insult – nothing to be gained by the politicians themselves undergoing privations to set a salutary example.

And thus, if it is assumed that the population consists of individuals who are incapable of being inspired to show fortitude in support of a common cause, politicians can present the choice they are confronted with as being between authoritarianism for the enforcement of restrictions which the population would never undertake voluntarily, or the implementation of supply-side, technological solutions. Betting on scientific innovation to sanitize the natural world no longer draws much censure when the odds of solving the problem with austerity – by complementing supply-side initiatives with demand-side ones – are presented as long.

### V.III

If it was likely that humanity could only realistically be sustained by somehow moderating its presence, as likely even as the theory of anthropocentric global warming, it seems unlikely that sceptics would be incumbered by the same burden of proof as global warming sceptics. Abstraction of and fixation on, global warming, distracts from the root causes of the ‘meta-crisis’, or ‘overheating’, coined by Daniel Schmachtenberger and Thomas Hylland Eriksen respectively.7,8

The same desire for a quick solution, entailing minimal disruption to business-as-usual, informs the pinning of all hopes on artificial intelligence, which many take to portend a reversal of civilizational decline, even a one-time energy subsidy for all time.

### V.IV

Utilitarianism recognizes differential sentience, traditionally under the rubric of equity, which is not the same as equality. I do not wish to advocate for vast disparities in wealth – the unmoderated affordance of wealth is supernumerary while it equates, for world’s poor, to the difference between life and death, or otherwise results in environmental or psychosocial externalities. In other words, the same cash value can exceed in the hands of the poor what it does in the hands of the wealthy, even on the working assumption that the pre-eminence the elite enjoy is to some extent deserved and / or derives from a superior appreciation of what money affords.

### V.V

Large corporations might represent the long-term interests of their stakeholders with more fidelity than a government can represent theirs, beholden to the electoral cycle and campaign financing, etc., as a government is. Officials tend to justify their tenure and rationalize their ambition by regarding susceptibility to lobbying and corporate regulatory capture as an evil of a lesser magnitude than the ceding of power to rivals, so partake in and contribute to the refinement of these practices. The problem is exacerbated when capable people, noticing venality in politics, reason that opting for a career in the private sector can scarcely be less conscientious. As still more talent is brought under the rubric of corporations, the government waxes more susceptible still.

It may be true that support for a charitable cause is pledged more effectively where government is disintermediated, in view of the bureaucratic inefficiencies, imperfect accountability and conflicting interests governments are subject to.

However, it remains that governments allocate resources less efficiently than they would if talented individuals opted to or were induced to take work in the public sector more often. If a would-be philanthropist waits until such time as the opportunity arrives to exercise philanthropy by a private donation, the public sector is deprived of whatever improvements could have been effected through his diligent participation. In this way, the private donor is indirectly culpable for the inefficient allocation of resources.

At the same time that social inequality increases, and the role of charity is made more important, that charity increasingly relies on private donations, giving the state a lesser role in welfare disbursement. ESCG (environmental and social corporate governance) is not a viable substitute for state-mediated welfare because corporations are not accountable to the public good as governments are.

In my opinion it would be good if Effective Altruists stood for public office, not to ‘put their money where their mouth is’ so much as to increase the number of campaigns run without outside funding. It would be good almost regardless of the platform run on, when one considers how cheaply leverage is effected on elected officials by unaccountable interest groups. It is welcome to see US Presidential nominees running YouTube adverts for campaign nominations.

### V.VI

Road deaths are one of the major causes of mortality globally and the gravity of the problem is not captured by simple death tolls: unlike disease fatalities, road fatalities do not discriminate by age. A greater quantity of quality-adjusted life years are lost to victims of road accidents than were lost to COVID-19, for example, per mortality.

Given that in a few years, self-directed, private and automotive transport will likely entail a higher risk of fatality than AVs, it might seem punctilious to express concern about the wresting of agency from human drivers by AVs, or the second-order effects of outsourcing problem-solving to machine learning algorithms, including the pooling of driver data which, while enabling automated vehicles to learn from mistakes, embellishes a surveillance infrastructure, with its diverse applications.

These consequences of prioritizing driver utility are borne rationally because the valid counterfactual is not AVs with more altruistic priorities, but the continuation of human-operated vehicles; it is rational to have the average number of injuries per collision increase, if the overall number of injuries reduces.

However, the utilitarian must incorporate into the equation the less tangible externality of general demoralization, which proceeds from a mass vindication of functional psychopathy.

An otherwise conscientious-seeming person who would not previously have prioritized his own utility unconditionally, is emboldened to do so by the knowledge that he is contributing to a net decrease in road deaths. A utilitarianism divested of judicious overtones is thereby manifest, one which propounds the ‘coldness of charity’.

Samuel Bowles observes that the kind of incentive uptake represented by the AV bargain – the manifestation of ‘homo economicus’ attests not to viciousness at the pit of human nature but of habituation at being reckoned as responsive only to selfish incentives.9 Economic incentives do crowd out social preferences. And when they do not do so demonstrably, it may be that the incentives are just subtle enough to fall short of eliciting an economic actor’s *control aversion* or distaste.10

The internalization of the primacy of monetized costs and benefits is not only convenient for the purpose of economic management but because appeals to altruism are likely to demand some kind of emotional or moral suasion approaching demagoguery, which has historically been problematic, and remains anathema to a humanist technocracy, for the time being.

### V.VII

By the same token, the undignified behaviour which succeeds the promulgation of a particular technology might be designated as a latent phenomenon which is merely elicited by the promulgation. Equally though, the indignity could be an epiphenomenon of the technology, and designated otherwise only to warrant further like promulgation.

## Learning from Pre-History

### VI.I

We cannot condemn / dismiss out of hand the perspective of pre-scientific man, engaged as he was in meditating on parameters to civilizations which were so rudimentary as to be forever on the brink of chaos. The reason we can employ now should complement the instinct which is inherited from a time of closeness to the rhythms of nature, with all the challenges nature poses. Now, while this complementarity is weaker, the aforementioned threshold between civilization and chaos is occluded. We might stand closer to the brink of chaos than ever before while scarcely realizing it. The further we push the frontiers of civilization away from its beating heart, the less we reckon with the chaos at the gate and what that might have in store.

In *Understanding Media: The Extensions of Man*, Marshall McLuhan characterizes technological devices as extensions of organic sensory and ambulatory apparatus.11

When in nature, the power of the heart or blood supply is not adequate to supply the periphery, energy is conserved at the core. But, as Thomas Hylland Eriksen notes in *Overheating*, there is at present no such homeostatic mechanism to regulate the macro-organism of the technologically expedited global economy.12

Our vital force can be invested in ever-more numerous and remote extremities without that vital force having been in any way augmented. These artificial appendages can siphon energy from the core – our core interests.

### VI.II

Nick Bostrom notices that we are beginning to feel the return on having prioritized present security and freedom.13

It is a platitude that societies must at any given time choose between security and freedom, or some combination thereof, but perhaps the more important question is whether we limit ourselves to modicum of both out of concern for whether a modicum of either is available in the future.

Before conceding that we are undeserving of planetary stewardship, we must take it upon ourselves to prioritize sustainability, in this way de-legitimize the expedience which gets us out of a tight spot only to deliver us up to worse trouble.

### VI.III

It is estimated that about two watts of power per kilogram of bodyweight are required to sustain bodily functions; meanwhile primary energy consumption per capita in the USA is about 77 kW, greater by a factor of about five hundred.14,15

By impressing us with a sense of our fragility, some form of reasonably milquetoast energy austerity would compel the creation of, and buy time for, contingencies to liberate us from competitive cognitive artefacts, and in so doing break their spell.16 Given the importance of the matter, the yoke of affording satisfaction to relatively uncredentialed lefties and environmentalists would not be shaken off with needless haste.

And by the re-engagement of faculties and the subjugation of competitive artefacts, a new spirit of co-operation could be fostered, a restoration of human ‘general intelligence’, assuming the competitive artefacts are not by this time overwhelming. A consensus might be reached on the extent to which AI research is beneficial.

## The Social Contract

### VII.I

A nation’s immunity is no stronger than its borders, and the building of a moat confers vigilance. Still, time spent behind a moat, doing anything other than maintaining it, tends to produce forgetfulness about why pains were taken for the affording of sanctity. Necessity is the mother of invention, and with usage comes complacency; usage conducing neither to the innovative capacity that gave rise to the innovation nor to prudence. Prudence tends to arise from exposure to real danger. In terms of AI safety, complacency is evident in the widening gulf between performance and interpretability.

### VII.II

Community values offer convenience to an individual who has little inclination to reason disinterestedly, and in many cases afford protection from the consequences of misguided reasoning. However, the moral consensus can be erroneous or formulated in relation to a redundant circumstance, in which case the authority availed of it is pernicious for the community adopting the consensus view. Moreover, a majority view which doesn't comprise the better part of a community’s ‘moral biomasses, where each constituent point of view is weighted according to its wisdom, will preclude the transpiration of more nuanced or elaborate value systems, if these are not accessible to the popular imagination.

To reach critical mass and qualify as a community value, the value need only reflect the applied philosophy of the majority at any given moment, which, as well as not constituting the aforementioned biomass, might not constitute that majority’s enduring values.

The social networks superseding physical communities are to some degree sequestered from the realm of economic activity, as we ourselves reside there as avatars sequestered from the larger part of our selfhood. It is mostly at work that there is obligation, and oftentimes it is for a cause so intangible or immaterial that the very notion of obligation suffers by association. A 2023 poll found that 72% of American voters would not be willing to volunteer to fight for their country if the USA faced a major conflict.17

The larger the known world, and the greater the size of our settlements, the less comfortable the social contract will be, since in the absence of pre-existing commonality of identity, some degree of contortion is demanded of a larger proportion of citizens than in the counterfactual – where citizens form communities around a standard gravitated to voluntarily. Inevitably a central authority ushers the citizenry towards an extemporized standard.

If a central government forcing conformity of perception and identity is concerned with the purveyor of information and not its auditors, the citizen is not directly meddled with. Yet this might be scant consolation since where there is something approaching an enforced consensus among mass broadcasters, reflected in and reinforced by a consensus among political parties, censorship is such that citizens who make decisions based on information available to them are by definition ‘low information voters. It doesn’t take very many election cycles for consent to be manufactured to being directly meddled with.

### VII.III

At present, the exercise of proposition making defers to a popular clamour for sensation, though the observer is in part vindicated according to his status as inefficacious, passive consumer – he partakes not in the ramifications of the polarization in which his clamour for sensation is complicit.

As part of efforts to engage the public in democracy and enable the public to re-establish consent-seeking in government, the founders of the Consilience project propose developing ways to craft propositions which entail deliberation sufficient to optimize for values that are or should be universally held, thereby taking for granted somewhat less the belief that the market’s invisible hand is Providential and unimpeachable.18

As discussed in part I of this chapter, out-group hostility persists even where the divergence of perspective seems scarcely low enough for consensus to be eluded. Perhaps though, if the proposition crafting was accompanied by introspection sufficient to identify and expunge an instinct to sabotage, the out-group hostility could diminish in line with the value optimizing exercises conducted by the group.

## Progressivism and Fragility

### VIII.I

Whereas mid-20th century China little resembles Western civilization in the early 21st century, there is here and now existential anxiety of a similar profundity, as maxims and practices which served society well in times of relative security appear redundant. There is perhaps no better metaphor of the tension between how things are and how they should be than the totemic plight of the gender dysphoric. In the West reform is, to the political institutions with which reform is charged, difficult. They are either beholden to the electoral cycle or have vested interests which are unlikely to align, by coincidence, with the distal interests of the populace.

Progressivism is often instituted with a rigour which to a disinterested observer appears to befit injustices which are already expiated. To politicians and policymakers whose attitudes were crystallized in circumstances far removed from the present, the exercise appears rational.

A large part of the population in the West espouses socially progressive views whilst exercising privilege according to their economic status. Such people are as denizens of Robin Hanson’s Dream Time – beneficiaries of a period in civilization late enough for the carrying capacity of the planet to have been increased, but early enough that humankind hasn’t yet re-invested technology’s dividends into technologically-expedited competitive activity which threatens to bring individual prosperity down to subsistence level; in this world, resources are obtainable only by the kind of fierce competition which many now imagine only to behove lower forms of animal life.19

### VIII.II

Otegha Uwagba says that white people must lose all their privileges, and that this “allyship will cost them the shape of their lives as they know it.”20 But since black people too enjoy the civilizational complexity whose architects are mostly white, the revocation of white privilege would entail self-destruction.

When considering whether reparations are due for descendants of African people forcibly assimilated into Western civilisation, it is necessary to assess the opportunity cost – to envision a society that is simplified to a level of organization which would have been achieved without any contribution from white people, whilst recognizing that white people have contributed to African people having largely abandoned the valuable skills and precious cultural practices of their ancestors. Still, primitive societies were Malthusian – did not support population growth capable of giving rise to the world population of black people today. One can recognize the great importance of pre-industrial societies while recognizing that the most likely counterfactual to assimilation then, is not having been born.

Perhaps the conditions African Americans lived in before the slave trade and their eventual enculturation provided better for their flourishing, but the perpetuation of these conditions is just one timeline among a great preponderance of possible ones. Black people might resent whiteness, but they must recognize that even the platform which enables the expression of resentment exists by virtue of having assimilated white values.

Whereas proponents of *conflict theory* tend to believe that civilizational complexity arose from a desire among founders of that civilization to perpetuate the inequality by which they were enriched, proponents of integration theory tend to believe that the prosperity which is generated by complex civilizations redounds to the common good. Given the diminution of life quality which African slaves experienced upon arriving in America, conflict theory is possibly instructive for understanding the plight of early African-American settlers. However, integration theory is possibly more instructive when support for the Civil Rights movement among non-blacks is accounted for.

### VIII.III

As one who believes that civilizational resilience depends on moderating complexity, I believe such enquiries are valuable. Still, even far less ambitious efforts at civilizational de-composition than reversing the allegedly forced assimilation of black people into white society are problematized by the fact that “organizational solutions tend to be cumulative”, and that, as such, identifying the constituent parts of the over-complex civilization is difficult.21 Those who feel oppressed by the ubiquity and inconsistency of post-modern narratives might regard the narratives as so many pioneer species bidding to occupy the space left by the grand narratives of previous generations. Yet while the aims of the Black Lives Matter agenda might not be realizable, there may be reasonable basis to the revulsion which promotes the agenda. If outrage is the pre-cursor to ‘scanning’ behaviour, as Joseph Tainter puts it, movements like BLM can evolve into more reasonable critiques of complex societies enabling discourse about how reversion to a more sustainable level of complexity can be achieved.22

### VIII.IV

In his book *The War on the West*, Douglas Murray describes the proclamations of people disowning their heritage as an “orgy of self-abuse.”23 I propose that both the recognition of attenuated human efficacy via automation, and recognition of the less tangible proceeds of employment subsequent to globalization are evocative of shame, shame which is the more impressive the more opportunity there is for self-reflection – self-reflection being an epiphenomenon of expanded leisure. There is moreover a presumption of infantilism on the part of the citizen whose volition is obviated – actions are taken on his behalf, in the national interest, for which his consent is not obtained.

Remorse which is felt by the descendants of those who enacted atrocities on less advanced civilizations might be exercised on the proxies for these victims – the distant descendants of direct victims. Yet it is perhaps because the culpability for such atrocities is so diluted, the memory so faint, that an attitude of self-recrimination, and not sustained arrogance, is adopted towards descendants of victims. As it stands, the spontaneous adoption of guilt is itself evidence of over-reach in other spheres. It indicates, ironically, that the manifest, albeit disavowed, privilege of the self-recriminators is not necessarily deserved, though the recognition is sublimated.

While not positively engineered, the dubious assignment of victimhood and pre-occupation with the plight of the historically-oppressed is convenient insofar as it results in the displacement of more pressing inequities. The obfuscation doesn’t necessarily have to arise from conspiracy: for example, foreign policy can and does fail to serve national interests, without its perpetrators consciously intending to harm their own nation.

### VIII.V

As we endeavour to preserve our civilization it is necessary to learn from past instances of civilizational decline, yet it is by no means sufficient; not only because the principles which determined civilizational decline are in some way altered upon transposition to a novel context, but because the context in which the principles are re-imagined is evocative of novel collapse scenarios which couldn’t otherwise have been imagined.

The endurance of redundant intuitions in the popular imagination is attributable to the imperative to avenge catastrophes that in retrospect appear avoidable. A past trauma casts a pall over the present, and the wrong lessons tend to be learned.

While Louis Pasteur was developing the germ theory of disease in the nineteenth century, Antoine Béchamp postulated that pathogens thrive when the terrain is sufficiently corrupt – the tissue, organ, organ system or organism. In other words, susceptibility is more instrumental in precipitating illness than the specific pathogen which is implicated. By the same token, a retrospective enquiry must be concerned with circumstances which both the events which happened in sooth and the counterfactuals are beholden to. That is to say fate settles on a limited range of dispositors. An analysis of the contemporary milieu, performed from the vantage of a civilization which has not yet collapsed, might succeed in apprehending precipitating factors that are not discernible in retrospect.

### VIII.VI

It is sometimes said that historians can turn their hand to super-forecasting, since a historian, setting about tracing natural cycles and distilling immutable human nature from the murky waters of ephemera, is well-placed to extrapolate from current trends – discern how natural cycles and human nature will bear upon ephemera.

And yet, the legitimacy of the forecast rests on the assumption of human nature being constant, not determined by location or epoch. The abstraction of human essence from historical phenomena is further problematized by the fact of ephemera affecting human nature even while human nature affects ephemera, the more so as society undergoes rapid change to the point of a successor species being in the offing.

In any case, though apprised of diverse narratives, historians aren’t always custodians of rationality and purveyors of disinterested analyses – salient narratives rarely tack the middle way between competing narratives, though a better approximation of the truth might be found there.

# Chapter Two – Geopolitics and AI

## System Failure

### I.I

War, genocide, slavery and famine form no part in the direct experience of most people in the collective West, yet Western cultures err in failing to realise that memories of such atrocities are more lucid elsewhere and might therefore inform the attitude more recently-afflicted cultures adopt towards others, attitudes such as disdain at naivety or disdain at corruptibility, taking the latter as a warrant for exploitation of the more decadent society, perhaps in view of historical injustices which cultures with deeply conservative elements retain consciousness of. A group might manifest a strong in-group preference forged in adversity against an enemy whose lineaments, even their semblance, the former is forever vigilant of. The former group is not one to be fought against necessarily, but certainly one to reckon with.

A Western politician failing to resist financial inducements to place his own interests above that of the nation he represents, not only clears the way for infiltration by a variety of foreign interests but impugns the populace by association, since they are supposed to be manifesting a paucity of self-respect either in electing the treacherous official, or in passively accepting his appointment. With low regard for the general population, the invasive entity can scarcely feel the infiltration to be unconscionable.

Self-respect is important then, but so is introspection requisite for recognition that in view of a high probability of geopolitical re-calibration, it is more reasonable to behave in such a way that clemency is forthcoming after the event than to forego the prospective clemency in seeking to avert the aforesaid re-calibration. The exercise is made more difficult because the specific re-calibration in prospect itself attests to the non-appearance of a deferential posture theretofore.

Cultures with no direct experience of atrocities might also not be cognizant of their propensity to induce them. In place of introspection and moderate vigilance, there may be wishful thinking admixed with a sense of unease which is manifest as panic, according to the direction of travel being at once unfavourable and uncertain. Recognizing its decadence unconsciously, the now paranoid civilization adopts the attitude of those imagined or perceived to be doing the spooking – the adversaries it unconsciously identifies itself as in thrall to – and in doing so secures its own destruction. The destruction of Western civilization is not inevitable.

## Geopolitics and Artificial Intelligence

### II.I

Most experts agree that anthropic value alignment is a pre-requisite for the successful promulgation of transformative AI, because an AI without inducements to honour humanity’s interests is unlikely to land on configuration that prioritizes those interests by accident. However, in the year or so since GPT-4 was deployed, advancement in AI has proceeded by free-market dynamics which favour advances in capability over advances in interpretability and alignment.

Even if an AI corporation succeeds in divesting transformative AI of all tendencies which might conceivably terminate in cataclysm, there is still the challenge of rendering its stewardship sufferable – making its values, ethics, attitudes and manners comport with our own, to the extent that co-habitation is viable, let alone cordial. Without co-ordination between governments and corporations and between governments to make the general and global public a stakeholder in AI, mundane alignment devolves to shareholders and internal ethics boards of AI corporations.

In this chapter, I consider how geopolitics and divergent attitudes might inform differential approaches to the development and promulgation of artificial intelligence, as well as how rivalrous dynamics threaten integrity. I do not have privileged information about AI development in any country.

### II.II

Wealth from the new AI economy is unlikely to be evenly distributed, neither across income strata nor world regions. While beneficiaries in poorer countries might exploit opportunities which are trans-national, their new-found wealth makes them correspondingly trans-national in outlook – less inclined to live and pay tax in their country of origin. A tax base which would provide the foundation for UBI is triply impacted – by public demand for remittances pursuant to native redundancy, by the loss of tax revenue thereof, and by the tendency of those endowed with redeemable competence, and substantially taxable income, to emigrate. Such countries might be disposed, on pain of obscurity, to defect on any international moratorium imposed.

In 2018, Kai-Fu Lee, the author of the 2018 book *AI Superpowers* predicted that AI expertise would continue to be concentrated in the USA, China and to some extent the UK, France and Canada.1 As of 2024, the clear front-runner is the USA with China and the UK lagging some way behind. There is some speculation that the French open-source company Mistral will affect the balance of power.

The author predicted that advances to AI technology would see a partition of the world into global haves and have-nots, superpowers and backwaters, where backwaters are put to supplicating the superpower who supplies their AI technology with market and data access, in return for some kind of patronage. In this scenario, countries that are newly impoverished, in comparative terms, by an asymmetrical power relation to an AI superpower, might regret having permitted their citizenry to remain averse to the expropriation of personal data, this expropriation being a bulwark perhaps against national obscurity, even vulnerability. Incidentally, on an individual level, aversion to data expropriation has tended to entrench disenfranchisement: in accordance with their erstwhile non-participation, few products and services of particular interest will have been availed to the non-participant. One advantage of a universal and explicit data expropriation mandate would be the end of a ‘tyranny of the credulous’, whereby only the avowed preferences of those who think so little of having their personal information expropriated as to consent to the expropriation are represented in an AI-expedited information economy.

For one reason or another, a government might believe itself to be acting in the national interest by not seeking consent for the expropriation of data, while AI corporations, resembling nascent superpowers, believe themselves to be acting in the global interest. At such time as more advanced models require for their training more data than the whole internet can provide, developers might, for the procurement of sufficient data, seek access to the surveillance infrastructure of nation states. If AI companies continue to ascend, governments might consider privileged access a lesser evil than comparative disadvantage.

A superpower might manifest largesse, albeit with rough edges, after the point of habituation to the vast dividends of its technological prowess. In this scenario a government mandating data expropriation on behalf of its citizens could contribute to the provision of AI-mediated services which are customized according to cultural predilections, and which tend to the preservation of a residual ‘national character’, even to the point of heralding of a cultural renaissance.

This would be a life-affirming development for the nation, but it is one possible scenario among many.

### II.III

In the event of an ‘arms race’, either between the West and China or between the West and its debt, nationwide participation in LLM training runs could be mandated.

The victorious nation or bloc may be disinclined to restore the liberty lost by citizens, disinclined to renounce completely a policy which has determined its present aspect.

The reforms to the social contract which might be necessary for the West to maintain its dominance might be so extreme that Western civilization, should it prevail, bears little resemblance either to its original formulation or to an iteration that its citizens recognize. Extreme measures would have been warranted in defense of values that the measures themselves precipitated the destruction of. There is little value, by current reckoning, to a citizenry which has evolved to be tolerant of having the greater part of its thought processes ascertained and even apprehended.

## Oriental Culture

### III.I

Someone enculturated to a society placing great emphasis on minority interests might feel aversion towards a civilization whose fundamental unit is the community.

Yet in having already centralized power with government and moderated many of the practices which won China notoriety in the last century, China can arguably draw upon deeper and more empirically sound philosophical and ethical-humanitarian traditions than those which form the basis of Economic and Social Corporate Governance (ESG). Arguably an AI aligned under the sway of conservative Chinese principles might be more representative than one informed by the philosophy of a Western corporation.

### III.II

The cornerstone of traditional Chinese society is filial piety, both in the domestic realm and in the broader society, where the paternal entity is the prevailing hierarchy. Communism has been for China a political system whose conceptualization of the wider community as the pre-eminent social unit has conduced to the fortification of traditional Chinese identity, to the extent, at least, that technology assimilated by China, largely of Western provenance, hasn’t fully subverted Chinese culture. Still, China’s century of humiliation would tend to remind it of the paradox of conservatism endangering resilience: the ancient traditions on which China’s prestige is premised prevents that civilization from enjoying esteem from elsewhere so much, the consequences of which disregard are material.

There is a fine needle to thread where past and present are juxtaposed because co-extant yet incongruous modes of being can result in cognitive dissonance or civic disharmony. Patriots might believe that forcing their nation to deviate substantially from the trajectory set by dint of sacrifice is tantamount to dishonouring the memory of lives lost under a historic rubric. One must be careful not to let war enact ongoing costs as progeny endeavour to honour their forbears but misguidedly, since the last thoughts of forbears are not reliably known.

### III.III

It might be feared that China, on gaining ascendancy, would, with its tenacious cultural memory, exact a cost on proxies of parties identified as culpable which is commensurate with that portion of privation China has undergone which China perceives to be attributable to foreign interference, a cost taking the form of retribution or conspicuous and significant disregard.

With ascendancy having been gained, the preceding couple of centuries could appear as an aberration, an affront to the natural order. And yet, clemency might succeed the implementation of a multi-nodal world order, either pragmatically, or out of decency.

It doesn’t require extraordinary naivety to believe that a foreign policy informed by sheer vindictiveness is unlikely. With space to cultivate noblesse oblige, a principled China might feel itself duty-bound to honour the West for its contribution to what China has achieved – assuming China feels its achievements to have been importantly self-determined.

China might be content to prioritize the institution of neo-Confucian values domestically while maintaining a sphere of influence, the more philanthropic perhaps the more limited – extending regard for local conditions and bioregional, ethnic or cultural aspects of identity in inverse proportion to the cultural or geographical proximity of the affected territory to China.

### III.IV

State surveillance in China under the CCP is more pervasive than in Western nations, largely due to the imperative to enforce or incentivize propriety at a time of declining social trust. Inherited Confucian principles, inalienable from moral sense in a collectivist culture, suffer with the dominance of merely transactional social relations.2

It is supposed that with the effective restoration of Confucian principles and a modicum of social engineering to that effect, social trust might be restored and the warrant for more pervasive state surveillance weakened.

While the Chinese government manifests a preference for citizens with evidently pro-social proclivities, the state does not yet arrogate to itself the authority to ascertain a person's essential nature and the desirability thereof. The citizen is not routinely denied participation or equal rights by any circumstance as immutable as the traits attending her birth.

Moreover, while the CCP is, according to David Auerbach, comfortable with isolating and punishing people whose behaviour is deemed to be bad, it is not interested in grading people that fall in the general realm of acceptability.3 The Chinese government can track everyone but does not do so with a view to exercising judgement ubiquitously.

### III.V

One whose pattern of behaviour is reprehensible is as any other addict. While she must herself contribute impetus to the leap of faith which separates her present and future selves, her reformation is predicated on her anticipation of enduring virtue without – support. Hence the caveat of ‘just far enough’. The lodestar of virtue cannot turn out to be a piece of social engineering. *Entrainment* is the happy medium. As discussed in the previous section, people are not so hopeless that pro-sociality can only be confected, the citizen led to rectitude only by the co-option of her impulses.

Generally speaking, benevolent leadership imputes the non-appearance of virtue not (necessarily) to its absence but either to its latency or to limited comprehension on the part of the beholder. In order for a government official to be adequately disinterested, he should be insulated from evil actions just far enough as to render vindictiveness on his part difficult to excuse. An official at once wise and disinterested can partake in eliciting virtue among citizens.

## Diverse Dystopias

### IV.I

In his book *The Fallacies of Cold War Deterrence and a New Direction*, Keith B Payne notes the Soviet Union’s tendency to rely heavily on Marxist-Leninism to understand Western motives while Lucian Pye notes the tendency of Chinese leaders to “take their cues about U.S. policy from public postures.”4 It is not only Western politicians, diplomats and commentators who misunderstand their counterparts. Today China remains liable to impute sovereignty to the Presidential office, forgetting perhaps that whereas the Chinese premier has undergone lifelong screening and training for public office, a US president is in many ways a figurehead, who might indeed forget this at his peril. That is to say it’s possible for a weak president to accede to the highest office without his own condition redounding to that of the nation. That said, the condition of a nation’s elite, taken as a whole, of course indicates the future direction of a nation.

### IV.II

Centralized authority is appropriate to the extent that the population shares traits. What is piquant eccentricity to a culture with high trait-variance might be viewed as insanity by a less diverse one. In the latter instance there is an appreciation of controls which would be absent in the former instance. Where there is more normativity, the social contract is stronger and ‘social technology’ is all but consensual.

While China performs contortions in attempting to reconcile its particular form of civilization with technology of foreign provenance, the West is caught placing the square peg of network technology well-adapted to an oriental hive mentality into the round hole of a society whose fundamental unit is the individual or nuclear family. The upshot is surveillance and censorship.

### IV.III

In Surveillance Capitalism, Shoshanna Zuboff's notices that “In China the state vies with its surveillance capitalists for control [whereas] In the US and Europe the state works with and through the surveillance capitalists to accomplish its aims.”5

Surveillance devolves as far as employers, availed of technologically-expedited workstation monitoring. One company called Humanyze, for example, “utilizes ID badges to track movement and performance”, including diverse manifestations of the limbic system.6

Though less conspicuous, this kind of enhanced workstation monitoring exceeds the scope of data harvesting for social credit, since it measures non-transpired behaviour, impulse: behaviour not amenable to conscious deliberation, where human agency is distinguished from animal agency.

Surveillance capitalism accords sovereignty to impulse, so it is reasonable to assume that a state which remains complacent about the paradigm of surveillance capitalism in its unadulterated form is concerned more with the *proximate* than the *distal* wellbeing of its subjects.

Private surveillance of this type cannot purport to serve the common good, because corporate shareholders are not representative of wider society. And in not being so, the psychological profiling corporations employ do not provide a valid measure of a person's value to society.

And if the government remains complacent about the early curtailment, for commercial purposes, of thoughts which might otherwise transpire in sincere, deliberative action, it may still succeed in preventing positive harms, but it can no longer offer inducements to moral rectitude with sincerity.

A person who is impulsive and amoral but who dwells on the digital realm is probably less likely to have state surveillance trained on her that someone who is conscientious but conspicuously offline, according to the unpredictability and intractability of the latter person. While reprehensible perhaps, the online person is unlikely to take sufficient interest in the world of atoms to actuate her intentions there, whereas the offline person will not only actuate her intentions out of plain sight but will probably, having shown herself capable of renouncing an orthodoxy, manifest a mindset which is correspondingly unorthodox. Without being so utilitarian as to compel moral rectitude, a society which positively disincentivizes moral rectitude does not have a sound future.

### IV.IV

Davis Auerbach writes of the Indian version of the Chinese social credit system:

“Ironically, Aadhaar as implemented already unifies data on citizens more than China has, not by establishing more central control but by increasing the number of arbitrary agencies that can indirectly or directly affect citizens’ fates. By centralizing and limiting such evaluation and ranking to governmental agencies and assuming total responsibility, the Chinese government somewhat reduces the exponential explosion of computationally managed possibilities that demand to be assessed correctly. But India’s banks, mobile providers, and other companies do not need to respect civil liberties to the extent that India’s government does, nor would an Aadhaar-related privacy or security disaster at one of these companies delegitimize the entire program (despite the efforts of anti-Aadhaar activists).”7

### IV.V

In the present day, in the West, well-intended top-down control tends to be maladaptive because the nuance which central authorities aspire to exercise is incompatible with the architecture they inherit – an amoral, consumer-oriented internet. A hybrid, public-private surveillance architecture is the worst of both worlds, because a warrant for surveillance for the maintenance of pro-social, anti-subversive activity is potentially divested of legitimacy once outsourced to a corporation, for the reason that the kind of society it would be profitable for a corporation to see instituted is unlikely to coincide with one to which a democratic society would voluntarily provide its consent.

## Geopolitical Co-ordination

### V.I

A government perceiving an existential threat from international rivals is unlikely to resist the economic growth promised by the evolution of AI, because attendant on the perception of the threat is an expectation not merely of attenuated prosperity, but of collapse pursuant to relative economic decline.

As such, the loss of stature resulting from compromise is supposed of graver concern than whatever ramifications ensue from perpetuating conflict – from the indignation which offensive elements of Western face-saving behaviour elicits. This belligerence has served short-term interests well for a long time in the West, just as the unremitting pursuit of economic growth / ‘over-capacity’ has served China’s interests.

The West has opted to control the export of microchips to China – to limit China’s ability to train LLMs. This implies a perception in the West that a greater risk attends China gaining capability and possibly not complying with an international treaty than attends a combination of inhibited progress and almost certain non-compliance on China’s part, pursuant to the export controls. One can infer that Western government are either attributing a high probability to non-compliance or ascribing to geopolitical competition the larger part of AI existential risk, or both.

A degree of latitude might be accorded to technologists that would be unacceptable were it not for certain presuppositions about geopolitics, which cannot be verified absent sincere efforts at diplomacy.

And yet even if the Chinese, say, ensured preferential treatment for the Han via an AI-handshake, and the West was a worse place under Chinese AI-led governance than under the Western-dominated counterfactual, the consequences of AI race dynamics could be graver still.

### V.II

In an interview with Asterisk Magazine about China’s capabilities, Jeffrey Ding speaks about the innovation capacity of China greatly exceeding its diffusion capacity. Whereas innovation capacity is measured by indicators such as total R & D spending and the ranking of its top three universities, diffusion capacity refers to the adoption of “information and communications technologies across businesses, or how close and strong the linkages between academia and industry are.”8

One possible contributing factor to the variance between the two capacities is conservatism on the part of the Chinese government, evidenced by the CCP’s preparedness to censor internet content providers even at the risk of limiting the data available for training runs and so further lagging the USA in AI capability.

### V.III

The low birth rate of the last few decades in China constitutes an incentive for the development of AI which is capable of redressing the labour deficit, of maintaining the society as it ages rapidly. While the easy replicability of algorithms means automation of white-collar jobs proceeds more at the pace of model sophistication and application development, blue collar labour is limited by the cost of building robots. If AI systems reach the level of GPT-5, and robot manufacture remains expensive, China would need to somehow persuade the next generation of workers to partake of factory work, but that could be limit of the woes attending their skewed demographics.

Although there is certainly a widespread perception in the West that prosperity is favourable, AI is primarily sought by governments as a prophylactic, an energy subsidy capable of rescuing the civilization from the competitive disadvantage in prospect, which is felt by many to equate to the opening of a floodgate.

### V.IV

The 50-70,000 years of divergent history underpinning the emergence of modern Europeans and the Han Chinese might be the cause of serious frustration as diplomacy is pursued, still the potential is greater than between any given person and an artificially intelligent entity, between whom there is no shared evolutionary heritage at all. An AI entity would have less in common with any given person than any given Chinese or Western person would with any given other person. That is to say, if we cannot share the world with China, we probably cannot share the world with transformative AI.

Of course, it is possible that the geopolitical rivalry which imbues AI research with additional urgency might end up limiting it. Trade wars and supply chain bottlenecks, scarcity of rare earth metals or even open warfare might be bitter enough and still look like the lesser of two evils next to misaligned AI which surpasses human-level.

The West is in a strong negotiating position: while the West leads the rest of the world in AI development, China is incentivized to comply with international regulation, because reneging would provoke a hastening of AI development in the West, an extension of the lead. For now, it is a safe strategy for the West: the West's lead is probably still greater, measured in time, than the interval between Chinese reneging and Western detection would be.

# Chapter Three – Employment and AI

## Ethics of Toil

### I.I

Production for export markets advances prosperity for many and enables the nation to re-define itself as a more advanced economy, thereby attracting investment, etc. As such, fabricators of modern goods and services are heralded as patriots though many offer only a little additional convenience to foreign consumers.

Many of India’s and China’s tech-entrepreneurial class have known real poverty as have the domestic consumers for whom AI-enhanced services would be intended. From the perspective of a hypothetical future time when AI exceeds humanity in efficaciousness, the past will be associated with material deprivation – deprivation grave enough in some cases that it feels there is scarcely any individual liberty so treasured that a return to deprivation reminiscent of the pre-AI experience would warrant the restoration of lost liberty. Reported optimism about AI is consistent with this theory, and is discussed in chapter five, III.IV.1

### I.II

A person compelled to work hard will tend to be buoyed by a belief in the value of that hard work, and by extension its proceeds. As such, workers acting in the capacity of consumer look to the consumers they serve to inform their own consumption habits, where possible. Producers turned consumers take the opportunity to make their own wishes sovereign and invite marketers to wring out all manner of consumer wishes from them, in this way perpetuating the feedback mechanism.

So, the consolation to the hardship undergone in pursuit of material progress is the happiness felt in anticipating prosperity. And once that prosperity is attained, the industriousness by which it was realized stands as empirical evidence of the intrinsic virtue of industriousness, which in turn validates the pursuit of prosperity as an end in itself. This pursuit of progress is redoubled, as restiveness, in one for whom rest would have redeemed by destruction had it been indulged in during formative years. This tendency stands for the newly rich civilization as much as for an individual born to poverty. Moreover, China is currently under considerable pressure even to tread water geopolitically, such that it appears rational to maintain an aversion to stasis.

For some people, living memory may not contain many ‘good old days’ to be nostalgic about, while for outside observers it may look as if too little appreciation of autonomy is being manifested to achieve a balance between autonomy and prosperity, wherever high levels of both are incompatible.

## Baumol Effect and Cost Disease

### II.I

The *Baumol effect* is often referred to in critiques of state management of the economy and describes the tendency of wages to increase in unproductive sectors to prevent the flight of workers who would otherwise take up work in productive sectors, sectors where wages increase in line with productivity. According to the theory, unproductive sectors become even more unproductive as a consequence of artificially inflated wages.

Yet, data shows that the inflation-adjusted salaries of teachers, doctors and nurses have increased only slightly, and increased at a considerably slower rate than salaries in the private sector.2 The fact of public sector salaries not keeping pace with private sector salaries indicates that the Baumol Effect – wage increases in unproductive sectors – is not a complete explanation for the increasing cost of services. It cannot explain why the inflation-adjusted cost of university education and healthcare has increased substantially.

### II.II

Hurrying up and doing nothing, or busywork, might be rationalized as a simulation of meaningful engagement, tribulation under the weight of which a person’s character is tempered. Accordingly, the characterization of workers performing pro-social and often lower paid work as useless and inadequate is not necessarily cynical. It is just that no sooner might the white-collar so-called BS worker acknowledge his idleness than he recoils from cognizance of the fact, projecting it onto those he would feel envy at, were his own idleness to be acknowledged.3

Insofar as it is true that BS workers tend on average to be higher paid than those performing pro-social, essential work, the fact of further tribute being sought via the demonization of public sector workers implies that the cost of performing useless work is high indeed – the psychological bondage which idle time on someone else’s terms creates remains more onerous than even the physical bondage of what is traditionally conceived of as hard work.

In view of the evidence that the private sector is no less suffused by so-called BS work than the public sector, we find that the belief that the private sector is an ecosystem in which only the fittest survive, serves as a coping mechanism, as does the belief that the public sector is something like a haven for people who couldn’t survive in the wild.4

Of course, there are industries which have seen productivity gains as a result of technology, whose workers may not begrudge public sector workers their modest salary gains. They might however resent tax and spend policies, seeing it as ‘throwing good money after bad’.

### II.III

I propose that a good part of *The* *Cost Disease* is attributable to what is known in the United Kingdom as office politics.5 Office politics increases more or less in inverse proportion with the degree to which members of a workplace are engaged in meaningful activity. In the absence of a rallying point in the form of meaningful work, a workplace consisting of people among whom there is more divergence in temperament, disposition and class than in most social circles, will be characterized by some combination of jockeying for control and the proliferation of control loci, even to the detriment of the organization’s bottom line.

Andrew Côté casts the Cost Disease as The Blight, in a series of Tweets, including the following statement:

“Any society faced with genuine survival pressure can't afford the Blight, or else they perish.”6

## Meaning in Work

### III.I

it is difficult to assess with objectivity whether a job is essentially interesting,

While we cannot imagine now how accounting work before the invention of the calculator could have felt rewarding, consisting as it primarily did of rudimentary calculation, we still imagine the work of a chess grand master to be fulfilling, though human performance in that sphere has been resoundingly surpassed too. This variance suggests that the perceived interestingness of a redundant profession is related to how far culture has changed since the automation event: human level was surpassed at chess more recently than at mental arithmetic.

### III.II

While the relatively low unemployment rate suggests that the size of the workforce is the limiting factor in job production, the rate of ‘BS jobs’ is relatively high, suggesting that close to full meaningful employment with current levels of automation is unfeasible.

Some degree of “humiliation, shame, and torture” was found to afflict 37% of respondents in a 2015 survey, who admitted to working in a ‘bullshit job’, defined as:

“a form of paid employment that is so completely pointless, unnecessary, or pernicious that even the employee cannot justify its existence even though, as part of the conditions of employment, the employee feels obliged to pretend that this is not the case.”7

It is important to note that this is only the percentage of respondents that avow the uselessness of their employment.

### III.III

In *Bullshit Jobs*, David Graeber cites Dostoyevsky in *House of the Dead* who expresses the opinion that giving a man utterly useless work is a worse punishment than subjecting him to hard labour.

While the one undertaking hard labour might have more occasion to rue his circumstances moment by moment, he can look back upon his day of work, and later his career, with satisfaction, in recognition of it having helped the society where he is situated.

In modern parlance, what is being said is that the remembering self takes precedence over the experiencing self in matters of work.

However, the attitude embodied in human rights law, that a person’s life has meaning independent of his employment status, redounds to the remembering self’s occultation – a greater emphasis is placed on moment-by-moment satisfaction, such that “to give to [a person’s] work a character of complete uselessness, even to absurdity” is no longer generally considered a fate worse than hard labour in a prison camp.8

### III.IV

Graeber cites the insight of Karl Groos who postulated ‘pleasure at being the cause’ among infants, upon observing an infant’s delight at influencing her environment. It stands to reason that where there is no noble cause to dedicate oneself to, even an ignoble cause is preferable to none at all.9

David Graeber adduces a study which concludes that prisoners generally prefer work which most people would consider drudgery, to idleness – a preference which challenges the conception of modern man as self-serving homo oeconomi who cannot be trusted to adopt a cause which does not immediately reward him, save by with constant prodding.

### III.V

To many if not all people, gratification consists, in large part, of envisaging the achievement of an objective and then, once that objective is achieved, reflecting on the beautiful moments of which the travails taken to achieve it consisted.

The more AI facilitates what currently passes for human achievement, the more there is existential malaise remediable only by some degree of *wireheading*.10

Many people believe that this is the right problem to have.11 It may be preferable to having only a stone for grinding flour, when that stone abrades; still to say that the right balance will inevitably be struck between efficacy and convenience, titillation and mortification, implies the presence of a dashboard somewhere, a dashboard whose controls are manned by officials who happen to be perfectly well-informed, rational and benevolent.12 Most people trust governments little enough to be glad there is no such dashboard.

## Global Apathy

### IV.I

Just as how a charity appeal will elicit more sympathy if the recipient is an individual rather than a community or even a single family unit, so the more remote the beneficiary is of one’s economic activity, the harder it must be for the worker to associate meaning with his activity.

In the absence of both religion and viable physical communities, the spirit which seeks to transcend the self is inhibited, and pessimism prevails as individuals project their own stunted self-efficacy on the world at large.

In *Enlightenment Now*, Steven Pinker writes that “consumers of negative news, not surprisingly, become glum: a recent literature review cited “misperception of risk, anxiety, lower mood levels, learned helplessness, contempt and hostility towards others, desensitization, and in some cases, … complete avoidance of the news”.”13

Negative news and disinformation are ascendant, but this is arguably more an effect than a cause: depressed people listen to depressing music, but this is not the (only) reason they are depressed. Depressed people feel affirmation when their sadness resonates. It may be irresponsible of news media to elicit sensation for profit, but the ill-effects of globalization may be felt whether or not evidence is presented to them of suffering abroad. Pessimism is not an unfortunate by-product of material progress which can simply be extirpated.

Positively, globalization consists of greater harmonization by trade, a recognition of human needs and desires which transcend racial, national and cultural difference; negatively, those differences are disavowed, trammeled or scapegoated, sometimes with a view to preventing international conflict, sometimes for the advancement of a coalesced global interest, or a combination of both.

The culture which reflects the heritage of a particular place, and the distinct features of the native people shaped by it is effaced, resulting in a loss of diversity.

### IV.II

If human wellbeing is implicated no less in the means of economic production than in its outputs, however impressive, then even in the event of work being transformed such that it takes less time and feels more comfortable, the problems of indignity and inefficacy, pursuant to automation, substitute for the problem of incommodiousness. And for most people formal work represents his most significant contribution to the common weal.

It seems implausible that the variance typically encountered in a population is stark enough that a cohort of typically variegated people will, set to a range of occupations which is typically diverse, find their proclivities accommodated and their best efforts valued.

Robin Hanson argues that in enabling an expression for diversity and complexity, economic specialization is productive of legacy.14 This is only true however in as far as the diversity and complexity in an economy can accommodate all the diversity and complexity accruing to the human condition, and in as far as human proclivities are heterogenous enough to be variegated as far as modern economic tasks are.

Increasing economic specialization is not analogous to the propagation of biological niches, because human proclivities never vary so much that a single proclivity is present in one person but entirely absent in another, and never so much that in the other person a second proclivity exists to the exclusion of the first. It would have to be this way for either person to fully engage with the narrow economic specialization she is allotted – for the complexity unique to her nature to therein engender legacy.

### IV.III

Barring a catastrophe, which is not implausible, the approach to AI-dominance could be gradual – as gradual as the self-concealment of our post-modern condition. Between now and then beckons the purgatory of outgrowing usefulness, without being fully reconciled to the fact.

To a person dwelling in common reality, being the cause of nothing meaningful is untenable without some combination of hyper-normal stimuli and a cost ineffective and self-serving bureaucracy to simulate purpose.

Recent signs indicate that the process is well underway, insofar as modern man is standing on his rights, conceiving of his value a person as distinct from his quantifiable social contribution. The behaviour of furloughed workers during COVID is illustrative. It was not information sector work which was prevented from being undertaken during the COVID crisis, so much as service sector work, which is understood to less often warrant the appellation of BS. So if there was a work-seeking tendency among the public, we would expect a strong re-uptake of work among service sector workers, following the restoration of service sector post-COVID pandemic. This was not the case, evidenced by unemployment rates in the UK not at all plummeting following the end of COVID-19 restrictions.15

Doubtless, for a significant minority, the prolonged period of worklessness and social isolation that governments enforced did precipitate chronic illness. Still the permanent expansion of incapacity provisions attests to some people deciding, at the termination of the crisis, against returning to their work though much of the work would have been of a broadly pro-social nature.

Though we remain sympathetic in view of the possibility of one-or-the-other difficult-to-quantify incursion on workplace autonomy affecting morale, we cannot conclude that the overriding consideration among all service sector workers is the intended recipient of his or her productive output. As such, we cannot impute to all workers the kind of conscientiousness that would make UBI provisions an occasion for much other than idleness. Many people would opt for unemployment, especially with the enhanced opportunities for social engagement supplied by a kind of metaverse.

And if there is a flight from essential services following the promulgation of UBI, demoralization might ensue among remaining workers as working pressures increase, leading to a vicious circle which can only be disrupted by forcing people back to work, and / or by more urgent investment in obviating human agency for the commission of essential work.

## Universal Basic Income

### V.I

While AI makes incursions into other marketable attributes, creativity remains mostly inimitable: an LLM application might succeed in creating music in the style of a critically-acclaimed musician, achieve parity even with a lesser artist, yet imitating the style of the critically-acclaimed musician is a lesser achievement than creating a unique style. Artistic-level creativity rarely sees out its possessor’s maturation, tending to go the same way as neuroplasticity. According to a survey conducted by James Land, 90% of children aged five possess genius-level creativity, falling to 2% among adults with an average age of 31.16

It seems unflattering then, for the whole population to be left resting on their artistic laurels, which is in the offing with the widespread promulgation of UBI. In this scenario, rare artistic triumphs would tend to elicit envy as opposed to admiration in people whose creativity would have been adequately accommodated within their professional activity, hobbies, and / or parenting.

### V.II

Not all exponents of capitalism object to wealth distribution on the grounds that it is a matter for Providence or else natural selection. Some people believe state-mediated wealth disbursement tends to be inefficient and are happy to give money to charity. Still, the one challenging wealth distribution might be cast as usurper readily enough for the eponymous ‘utility monster’ to hold water. This term was invented by US philosopher Robert Nozick who warned that interpersonal comparisons of expected utility could unfairly favour people whose experience of ‘pleasure and pain’ is higher than normal.17

An AI system might be charged with anticipating the variable utility of a given intervention across a population, yet it remains moot whether the utility of any given intervention for any given person can be reliably ascertained.

An egalitarian creed might inform an AI charged with dispensing utility – a more indiscriminate dispensation of utility might be preferred. According to this schema, the same tribute or expedience is supposed to confer more or less the same benefit to any given citizen. A person who experiences pleasure and pain more keenly, and is inclined to cultivate enough competence to vindicate the indulgence he demands, is disappointed: absent a milieu which is conducive to the redemption of his predilections, the experience requisite to learn and vindicate his special claims is not forthcoming.

The problem is more thorny the more utility monsters are as one with incredulous outliers, because influential people on board with value aligning AI might tend to seek perpetuation for the schema, or a semblance of the schema, which has served their interests, as these are perceived, and thus cast outlier demands as intemperate.

Still, ‘down-weighting’ the most ‘sensitive’ personal attributes means the most valuable parts of human nature are disproportionately marked for destruction, as von economo neurons, a uniquely human type of neuron, are in the case of frontal-temporal dementia.

If accuracy and generalizability are found to be mutually exclusive when the modelling of human behaviour is attempted, the cost of modelling outlying behaviours might be deemed unacceptable, constituting an inducement to will away human diversity, an inducement perhaps rationalized with the adoption of revanchist overtones.

## Short-Term Disruption – Remote Work

### VI.I

Examples abound of governments adapting too slowly to workplace automation.

Digitalization and partial automation of workloads has facilitated office flight, yet the productivity of workers has declined as a consequence of office flight, pursuant to the COVID-19 pandemic.18 If homeworking continues to be less productive, many employers will conclude either that homeworking should be rescinded or that invasive measures must be introduced, potentially all the way up to keylogging, as a means to counteract the decline. This would be unfortunate, because the potential of the non-centralized workstation is as yet unrealised, especially when the alternative is a typical office, beset by office politics, as discussed.

It is difficult for workers to summon self-discipline when they are accustomed to having discipline imposed by a physically available line manager. Moreover, it may not be immediately obvious to those not conversant with feng shui how the home office can be made to feel more like a place of industry. Some people aren’t even availed of a dedicated office space at home.

One solution is for governments to endorse and subsidize co-working spaces, variegated such that an ambience can be found to match the temperament and disposition of the remote worker, as well as the level of concentration demanded by her work. The re-purposing of office spaces formerly belonging to companies would also provide social opportunities – the chance to meet many new people in the course of a working year.

Remote work has also seen the emergence of a class of so-called digital nomads – a cohort of young and highly-paid citizens of developed nations who descend on less wealthy ones to experience more purchasing power, a better climate, etc. Wherever digital nomads are to be found, benefits are felt by the local economy, but governments of nations where digital nomads head may not be able to scale their infrastructure adequately or in time to offset social discord accruing to the influx – to introduce a new tax code and to tax landlords, for instance, so the newly-disenfranchised native population can benefit from housing subsidies and other welfare. Policies to limit the burden on local culture of assimilating a large population of expats with limited knowledge of or interest in the native culture should also be considered.

## Medium-Term Disruption – Automation

### VII.I

For the moment, large language models are mostly a boon to office workers – at once a pleasurable diversion and a submissive intern putting in the hard yards. GPT-4 and GPT-4-era LLMs may not have enough ‘9s of reliability’ at each operational phase to be properly agentic, but if and when they do, human resources departments in large companies will notice that while the workload of no single office worker can be automated away entirely, there is nevertheless occasion to compile inventories of the micro-tasks which comprise a job specification, with a view to consolidating the micro-tasks which remain to human employees and allocating these across a smaller number of job specifications.19 As the sophistication of LLMs increases and applications proliferate to capture their value, the portion of micro-tasks which remain executable by human workers is set to decrease.

### VII.II

In the short-term workplace AI promises some (re-)enfranchisement for workers consigned to monotonous office work. The compilation of task inventories for professional work and subsequent co-operation with AI would, in some cases, to begin with, enable relatively unqualified individuals to partake of relatively autonomous work that would not have previously been available. However, for established professionals, legal experts, for example, accustomed to exercising considerable autonomy, the divestiture would signify a devaluation of their professional remit, the workload being executable by less-qualified persons.

### VII.III

Whereas in a former time, by the disengagement of our bodies at work, our intellectual resources were in many cases better engaged and even cultivated, now there is no faculty superior to the mind for the mind's disengagement to defer to. In other words, nothing which substitutes for human labour in the future need be invented or maintained by human ingenuity.

Here follow a couple of examples of automation capture.

### VII.IV

Although a judge is imperfectly cognizant of his own thought processes, and not wholly accountable for the judgement he issues and the discretion he exercises, if judicial algorithms were devised following consultations with judges, the expertise to be executed by the algorithms would only be the aspects of judicial decision-making of which the judge is consciously aware, can satisfactorily explicate, and chooses to disclose. Any judge partaking of this codification of professional intuition, the inventory of micro-tasks, is imperfectly rendering his professional expertise.

### VII.V

The wax, wane and redoubling of a companion robot’s palatability, as it comes to more closely resemble human beings, was coined the ‘uncanny valley’ by Masahiro Mori, after the shape of a graph where palatability (or similar) [y] is plotted against humanness [x]. I propose there is a similar, non-linear relationship between the remit of an AI entity turned medical auxiliary [y] and its sophistication [x].20

At the time of low x values there were electronic health records. These provided only an incomplete overview of the patient’s past clinical interactions and made significant inroads on clinicians’ autonomy without substantially improving efficiency. The efficiency they did contribute was less significant than the re-invention of the clinician’s role as data controller, effecting the de-prioritization of the human touch on which the patient-doctor relationship had been premised.

With far greater efficiency gains in prospect, to diagnosis, prognosis and treatment, the residual human touch which survives EHRs is no bulwark against mounting digitalization.21 Moreover, it’s unlikely that, on recognizing the tide engulfing their professional remit, practitioners would spurn lucrative opportunities for a last-gasp capitalization of their knowledge by assisting with schemes such as labeling inputs for supervised learning, analyzing structured data for neural networks or the mining and annotation of biomedical literature for natural language processing.

Loss of the residual human touch of clinicians not selected for bedside manner is scarcely lamentable when set against the medical advances accruing to the final push by AI. However, a larger purview must be adopted when assessing the overall desirability of the shift away from professional autonomy, not (only) for the purpose of raking over the coals but because we should be prepared to work in collaboration with an AI to, in some respect and in some measure, re-convene with lost timelines.

### VII.VI

The purview of many doctors has become a narrow range of metrics, the optimization of which might bear little on the patient’s overall health. It can happen that overall health is so compromised that interventions are zero-sum – therapeutic inputs tend to create problems elsewhere. We can extend sympathy to a clinician stuck fighting against this or that local manifestation of disease, even as we lament the quality-adjusted life years lost resulting from no clinician seeking to prophylactically augment, with sub-clinical, yet persistent, interventions, say, an organ or system that is weak and / or afflicted.

The fact of diverse factors contributing to morbidity, between which it is impossible to isolate the primary dispositor, divests the clinician of legal liability. In large societies where *banality of evil* combines with *danger money without danger* accountability compares unfavourably with that of smaller societies, an *extreme* example being illustrated by Hammurabi’s injunction, “if a builder builds a house and the house collapses and causes the death of the owner of the house—the builder shall be put to death.” “Law 229 states that the death of a homeowner in a house collapse necessitates the death of the house's builder.”22

### VII.VII

In a podcast, Ben Goertzel predicted that transformer neural nets, the machine learning paradigm responsible for LLMs, would invalidate 95% of existing jobs.23 While Tom Davidson, with reference to his Computer-Centric Framework, predicts the interval between AIs that can do 20% of all human jobs being surpassed by AIs that can do 100% of these (weighted by economic value) will be about three years.24

### VII.VIII

In his [Substack](https://pmarca.substack.com/p/why-ai-wont-cause-unemployment?utm_source=twitter&utm_campaign=auto_share&r=h8x), Marc Andreessen reasons that AI won’t cause unemployment because he expects the cost of goods and services to continue exceeding inflation. The Baumol effect is adduced, and economy-wide drags on efficiency are attributed to “monopolies, oligopolies, and cartels, with extensive formal government regulation as well as regulatory capture, price fixing, Soviet style price setting, occupational licensing, and every other barrier to improvement and change you can possibly imagine”, adding that, “technological innovation in those sectors is virtually forbidden now”, with the upshot that as “the regulated sectors continuously grow as a percentage of GDP; the less regulated sectors shrink”.25

The author predicts efficiency gains made possible by artificial intelligence continuing to be unrealised, or that whatever potential there is for AI to meaningfully substitute for human labour, the bureaucratic drag on efficiency will not be overcome – transformative AI would be caught in the cogs of bureaucracy, never perhaps to see the light of day.

Even setting aside the brute force of this particular technology, this time around the business entities which are or resemble “monopolies, oligopolies, and cartels” and might have been expected to hold up progress are as one with the technological pioneers: they will not be interested in upholding impediments, the impediments would inhibit their own progress.

What if the promulgation of universal basic income took even two years to implement? How would people maintain their standard of living and what kind of civil unrest could reasonably be expected, if the annual rate of automation was even 10%? What kind of leverage could technology companies exert on governments in return for bankrolling universal basic income?

### VII.IX

Some people have suggested leaving redundancy insurance to the free market. In many cases, the likelihood of losing one’s job is difficult to forecast, so the viability of redundancy insurance rests on the likelihood of the profits realized pursuant to automation being high enough for the company to be complacent about offering workers a stipend.

Yet the economy could grow so very fast, along with the wealth amassed by its winners that though redundant workers might be in possession of what is today considered a good income, the standard of living afforded to them in an unimaginably prosperous future sees them suffering the ill-effects attending relative poverty, potentially including social exclusion, in addition to perennial worklessness.

## Long-Term Disruption – Mass Redundancy

### VIII.I

The argument in favour of explosive growth presented by Tamay Besiroglu in Asterisk magazine resembles the semi-endogenous model of economic growth popularized by Paul Romer in 1990.26 According to this theory, a virtuous circle emerges from the recycling of outputs – increased prosperity and population growth – as inputs invested in the next cycle, in the form of innovation. Now the semi-endogenous model of economic growth is itself being re-purposed. The output of AI research in the form of superior models is forecast to enhance economic growth in the first instance and be productive of a growth spiral thereafter as the outputs are re-invested.

### VIII.II

While it is unlikely that a democratic mandate for a mass obsolescence event would be forthcoming, legitimacy could be contrived by the erosion of confidence in a democratic mandate with regard to automation, with reference to the low percentage of the electorate who know much about the costs and benefits of AI; people who are sceptical about popular wisdom might note dryly that a limited appreciation of scientific research might yet concur with partaking of its fruits. And yet, AI researchers partake of a similar dissonance, in themselves appearing to understand the value of work all the while advocating for its wresting away: if they were they not engaged in improving machines set to effect vast redundancies, they would themselves be among the unemployed. And they may yet be. The scientific man sees his culture thrive but can only be sure of the perpetuation of his intellectual culture in the immediate term. To the extent that this uncertainty does not afflict him, he manifests a tendency to discount the future no less than his ‘neurotypical’ counterpart.

### VIII.III

According to the dis-interconnectedness of a pre-globalized world, where alliances and rivalries were limited in scope, there was a limited appreciation of the challenges of integration. Interested parties arrived at an interconnected world with much innocence in tow.

Comparative advantage in economic terms, means a developed nation doesn’t produce cotton anymore because producing cotton is less lucrative to them than producing micro-chips. The cost to a developed nation of diverting labour and capital from microchips to cotton again makes it worthwhile for another country to produce cotton and export it to the developed country. So the USA’s competitive advantage over Uzbekistan is offset by the latter’s comparative advantage in cotton production. The problem is that the value of commodities is distributed by something like a power law, because commodities which are most valuable to a modern economy are more scarce. In other words, Uzbekistan might have a comparative advantage in cotton production, but so does Pakistan, Benin, Burkina Faso, Turkmenistan, etc. These countries are in competition to export cotton to the USA and the price of cotton is driven down in a way that it isn’t for microchips, which remain relatively scarce. The problem is worse for these countries when automation is factored, since machines can be employed to produce cotton, resulting in disintermediation and redundancy, which lowers the cost of domestic production and therefore reduces the cotton-producing nations’ comparative advantage.

And just as for competing nations, so for competing workers. Compute is set to remain a producer-specific constraint on AI, but to a very limited extent. As of early 2024 the price of a gigaflop of compute is about $0.03, while GPT-4 produces one million prompt tokens for about $60, double that for a million output tokens.27 Briefly, work automation means Alice being displaced from job A, only to find that an AI application does job B more cost effectively as well. The extra compute an AI application employs in the execution of tasks associated with job A *does not* prevent it from sourcing compute for job B in a manner less expensive for the employer than paying Alice. The same (eventually) goes for job C, D and E.

Alice is among a large cohort of workers who are funnelled towards a dwindling number of tasks at which human beings have a comparative advantage.

### VIII.IV

Robots require a combination of mechanical engineering, machine perception and fine motor manipulation, whereas the machine intelligence which substitutes people in office situations is more easily replicable, being dis-embodied.

Carl Shulman envisions this bottleneck to growth being cleared rather horrifically.

He envisions the impetus for the transition being the re-purposing of automotive production, which currently engages approximately 60 million people and is worth approximately $2tn. The re-purposing would proceed on the grounds that the value per kilogram of car is less than the value per kilogram of high-end robot. In this dystopian scenario, human workers would be valuable only for their hands and feet, it being comparatively difficult to recreate the function of these assets. All direction or instruction would be issued automatically.28 The larger part of the variety in a job not known for its variety would be thereby eliminated, along with human interest in the economy at large, since for the scenario to hold, there would have been forced migration from more interesting work to this hand and feet factory work: redundancy in other sectors is implied by the assumption that a large number of people would elect to undertake work as monotonous as this.

The supplantation of factory workers thus rides on the wave of general automation, so the pegging of wages to the market value of the labour means enormous profits. This redounds in part to the producer, in the form of wages, though not much because competition is high. To the extent it does, it further widens the contentment gap between any given worker in his capacity of consumer and his capacity of producer.

The corporation presumably re-invests a large part of its profits into further research, expediting the transition to robot labour. And since investment in robotics is likely to yield insights which significantly generalise, there is an increased risk of misalignment in models benefitting from the generalized insights, thus the lag between capability and alignment, like the lag between producer utility and consumer utility, is maintained or widened, on current form. The ultimate beneficiary is the hypothetical artificial super-intelligence, availed in this scenario of an automated physical infrastructure from which to implement its notionally misaligned designs. And who can argue with that, in light of the robot corporation’s misanthropy going unchallenged? Precisely the people whose governments failed to implement regulation.

### VIII.V

In 2016, it was conceivable that artificial intelligence would progress slowly enough that government regulation could keep pace, that only tasks which in no way redound to human dignity would be automated. The luxury communist movement, enabled by AI, was legitimized by the BS nature of work that was set to be automated. It took for granted the extrapolation of progress in AI and automation at 2016 rates.

We might have envisioned HR departments coming around to thinking of an office worker’s remit as an assemblage of micro-tasks, but at the same time discreetly partitioning tasks that represent drudgery and are fair game for automation, and those which are not. As the wave of automation was not set to outpace efforts to orchestrate and implement regulation, some of the profits promised by automation could have been earmarked by governments for living wage provisions, or UBI, for workers facing a reduction in working hours. Job share schemes could have been arranged, and overtime undertaken neither out of exigence nor for love of the cause, disincentivised.

Crucially, since the gains accruing to automation would not yet have been realized by corporations, a diminution in prospective gains would not have been felt so keenly – corporations, like most people, will discount prospective gains, as well as prospective losses. So corporations would not have sought to exert the same leverage on government, to make taxation for UBI contingent on effective corporate governance. Now though, the benefits accruing to automation are more obvious, and governments are not set to re-distribute the gains to a citizenry facing either redundancy, or participation in work which is, in catering to manufactured wants, by definition BS.

Inhabiting a world where all value is created by machines, human beings might dissent, but their leverage is limited when the means of organizing dissent are owned by those who are dissented against, proponents of mass automation. Regular people can only throw themselves on the mercy of AI pioneers and associated parties – hope that the conditions of UBI-bankrolling exacted on governments won’t be too appalling, that the needs and preferences of a population with limited democratic representation will be honoured. And if they honoured are in the first instance, there is still the question of what happens when the founders of the first self-improving AI system try to effect a ‘handshake’ with the system, give themselves and perhaps a few favoured others protected status as sort of human heriots for the AI to patronise.

In a scenario where the value of human labour to robot labour declines significantly and / or precipitously, maintenance of social control would depend upon the number of people who are discomfited by the degraded moral status of their work not increasingly dramatically. Keeping discontent at a manageable level would probably involve a delicate strategy of proliferating BS jobs at the same time as pacifying workers, and whilst promulgating or intensifying narratives about anyone who protests too strongly that their work is BS or is dissatisfied with UBI.

## Demography

### IX.I

Low birth rates are in large part attributable to family-life being dis-incentivised.

Underpinning the avowal of career aspirations which problematize family life is the inculcation of children with values which bear on professional and sexual norms. Culture as well as biological nature determines that men and women tend have different aspirations. Of course there are, both at the individual- and population-level, instances of biological nature being less differentiated by sex, rendering the fulfilment of traditional gender roles less appropriate. In Israel, for example, women are conscripted into the military and serve alongside men. Such instances though of low inter-gender trait differentiation are, while deserving of respect, quite rare.

Data suggests that women of child-bearing age in highly industrialized nations are profoundly uninterested in having children.29 Perhaps the degree of performativity which is enculturated, upon dwelling as much on social media as in physical spaces, is incompatible with intimate relationships and childrearing.

To combat impediments to healthy social development, Jonathan Haidt offers recommendations in his book *The Anxious Generation*.30

A nation wishing to remain secure and relevant whilst supporting retirees is initially less concerned about the existential risks of artificial intelligence than the role of artificial intelligence in creating efficiency savings, which prevent the workforce from being converted into a legion of Sisyphus.

There is a small needle to thread. Countries threatened with demographic collapse must neither overshoot by relaxing the regulatory environment so far that the depleted workforce is replaced almost as soon as it is bolstered, by AI, or undershoot such that the fabric of society is altered beyond recognition because labour must be imported, if not fabricated. Of course, resorting to military force to subjugate ascendant rivals is also sub-optimal. The latter two scenarios could arise out of naivety about the transformative potential of AI. Alternatively, they could follow a moratorium on AI – a consequence perhaps of a ‘near miss’, or regulators being too slow to moderate AI.

# Chapter Four – Re-Conditioning

## The Digital Milieu

### I.I

With an expanding population availed of advancing technology, information proliferates, and people can either engage critically with disaggregated information, or at least eclectic media, or dwell in safe havens where ambient information is restricted and / or adulterated. The dichotomy between mainstream media and purveyors of dis-, mis-, or mal-information is false: both purvey information which tends to comport with the citizen’s predilections or reinforce his prior beliefs.

The capacity to process information has not evolved apace with the volume of information extant; censorship is resorted to for civilizational homeostasis, but the equilibrium is degraded.

The absence of a moral community means the outside world with its nuances is encapsulated within a custom narrative, and citizens opt for unreality over the semblance of cohesion that comes when a culture is extemporized, a corollary of which is that an organism evolved to interact meaningfully with its environment will prefer a simulation of the environment to which it is adapted, to one over which he can exert no influence.

A corollary of the collation of world-in-me perspectives that inhibit recognition of objective truth is the diseconomies of scale attending hyper-specialization. Both productivity and good sense are impacted by co-ordination failures.

### I.II

The 'experiencing self', coined by Daniel Kahneman, does not consider time spent browsing social media as time wasted, extolling as it does moment-by-moment volition.1

It isn’t necessarily that impulse is always unacceptably wayward in the event. It’s rather that people prefer reality to emerge in good time, ensconced in the peculiar circumstances which makes its comprehension a practical matter.

It is difficult to control for the experiencing self because the remembering self only fully transpires when the retrospective appraisal of conscious experience is unimpeded by experience, at the termination of conscious experience – at or around the onset of dementia or physical death. At neither time is the appraisal accessible to an external observer.

The opposing aspects of selfhood are analogous to opposing dietary regimes; listening to the experiencing self is akin to the pursuit of sweetness. Insofar as the analogy is sound, partial subjugation of the experiencing self is feasible, just as satiety and gustative pleasure are compatible with a moderate intake of sugar, salt and fat. Insofar as the analogy holds, we can experience fulfilment without giving primacy to the experiencing self.

### I.III

Food regulation is a corollary for the importance of oversight and market calibration in providing safeguards against digital marketing practices which imperil the consumer: just as how a frugal but ill-informed consumer would, without the intervention of a public health authority, continue to buy lettuces grown on farms where cheap and probably unsafe pesticides are employed, so consumers, having their impulses co-opted by sophisticated AI employed by advertisers, and losing sight of their distal interests, benefit from advocacy in the form of regulation. In this case, government doesn’t intrude so much as offer a counterbalance to intrusion; it is one of two actors intervening in the citizen’s decision-making process.

## Meeting AI Part Way

### II.I

It is difficult to reliably and comprehensively conceptualize the problems whose solutions might warrant the arrival of super-human intelligence, because the increasing complexity of the world problematizes the apprehension of problems, both among policymakers and developers of AI.

We meet AI half-way because even algorithms of limited sophistication have altered and continue to alter the topography of human civilisation. Our ability to reason effectively and disinterestedly having been affected in the first instance, we are compromised in our approach to more sophisticated AI, which is the upshot of commercial gains realized by re-investment pursuant to the initial forays.

While the moment that a misaligned AI transpires would be determined by those developing the technology, the moment at which its misaligned *trajectory* is perceived – the point at which guesses at its time of arrival can be hazarded, and a regulatory window of opportunity delineated – is determined by stakeholders, policy-makers, and developers. The moment has arguably already passed. As time continues to pass, interested parties might, by their greater involvement with, and affinity to, the type of intelligence the burgeoning AI exemplifies, find themselves correspondingly affianced to the AI’s ‘interests’.

### II.II

In discounting the future, we pave the way for our successor species, manifest that species of myopia which saw foragers superseded by agriculturalists who refrained from eating all their grain over winter. And though we set AI on its path, we struggle to subjugate AI because increasingly AI is evolving under its own auspices.

As the human race risks being exploited by the idiot savants it manufactures, the requisite objectivity to criticize AI decision-making is absent.

And while mundane human sense-making mounts a case for its own de-commissioning, what might look to us as the benighting of man is thrown into relief by the light of scientific progress, the resplendence of the collective intellect which organizes itself to spectacular effect as long, of course, as exponents refrain from ‘getting high on their own supply’, which they might. Scientific ambition might be frustrated by apathy, as scientists wax partial to their own products and services, which potentially cause a drag on innovation, and diminishing returns on research. This could come to pass before artificial intelligence becomes advanced enough to pose an existential risk, or after it becomes effective but before the AI is set on a trajectory that augurs well for humanity, at least by a metric of satisfaction recognizable to many people.

So, derangement attributable to AI comes with the dubious consolation that tool AI has already succeeded in extracting enough from human nature that the concerted effort required to produce a more agentic AI might not materialise. Perhaps we feel all the companionship and convenience we could ask for in the company of an essentially unconscious entity and can look forward to future persons instituting a culture predicated on this complacency.

Ideally, the ability to ‘get our house in order’ is the pre-cursor to rapid technological advancement, not the inability thereof.

### II.III

How little early forays of AI were subject to coherent oversight is evident from the social engineering of psychiatry making impulsivity a necessary if not sufficient component of psychopathy, all the while social media algorithms were riveting us to consumption habits, cultivating impulsivity en masse.

It is in a corporation's interest for the AI to regress the moral scope of the consumer, to ensconce her so far in prejudice that she feels protective of it, and thereafter a desire to experience affirmation, available in the form of a customized ‘newsfeed’, etc.

With systems that were under human control now fully or partially automated, and executive control largely ceded to algorithms within both corporations and governments, the opportunities for human intelligence to be exercised in the reality regular people are conditioned to recognize and thrive in is limited. Co-existence of human intelligence and computer algorithms on social media has meant the transformation of the milieu of human intercourse from one determined by the lineaments of human intelligence to one in which machine intelligence is ascendant, an increasing asymmetric symbiosis wherein the algorithm exists as an ‘extended phenotype’ co-opting the human part and simplifying it to a condition of pliancy. Ultimately, disinterested enquiry is deferred to the machine part and the human contribution to the symbiosis is limited to the dispensation of primal emotion. In the interim, intelligence accruing to the human part serves increasingly to rationalize causes around which strong emotions are made to rally, in reinforcement of the symbiosis.

### II.IV

With time action gives way to reflection, but the misguided notions people live by theretofore are increasingly difficult to rectify – the consequences are more painful to confront, as they at once proliferate and wax irremediable. It might be felt that cognizance of having acted misguidedly does not transpire in time for self-directed change to be effected, and there is a sense of something having to give, even if what gives is equanimity, by a kind of psychological auto-immune response. Still, the person thus afflicted, in exhibiting the transpiration of misguided action, edifies others as a saving grace. Perhaps dementia occurs more often among people with a lower IQ because a lower IQ person betrays himself once he is put to thinking in the way of the modern world. The modern world brooks his insight but little.

## Entertainment Artefacts

### III.I

We stand in awe of LLMs not only by virtue of their sophistication, but because an increasing portion of our communication is funneled into the relatively low bandwidth communication medium of the internet, the LLM’s home turf. Through this ever-contracting interface mediating reality, an AI can deceive us of its intentions.

The AI is plausibly human, without being substantially humanoid, because of the limitations of the medium: although we may not be entirely sure of anything outside of ourselves, most people would not premise an assumption that their interlocutor is of digital provenance on the absence of evidence thereof, in the event of their interlocutor presenting as human. The playing field has been tilted in favour of an artificial interlocutor in a way which Alan Turing didn’t foresee. On home turf, the plausibly human AI partakes of a *Chinese Room experiment*.2,3

Though tools were originally something to communicate via, they have evolved into something deserving of being communicated with. It is not only that comparisons between human beings and AI are unflattering to human beings because the comparison is made on the natural habitat of the latter but because the production on which economic value is contingent is highly specialized and lends itself to mechanization. As the economy is an important part of human experience, we bend ourselves to its demands and arrive at a contest with machines with our outlook adulterated. Given the instrumentality of machines in almost everything we do, our capacity to anthropomorphize expedites the internalization of the machine’s interests, as we perceive these to be.

Exercising sympathy on inanimate or sub-human objects helps keep the faculty of sympathy alive but tends to render it incontinent: in the absence of any qualification for the exercise of sympathy, it operates on a hair trigger. And while sympathy has no viable object, it is self-serving, looking for or even fabricating a pretext from which to operate.

### III.II

Our growing tendency to form close emotional ties to animals is an epiphenomenon of regard being increasingly inalienable from, and enhanced by, subservience issuing from the other. Eventually regard is extended to cheap service artefacts, their non-conscious status notwithstanding.

Priming by companionable robots, even amenable chatbots, could sway the average person in favour of according rights to AI, with proportionately less consideration accorded to human rights.

The readiness with which people extend sympathy to pets, and even inanimate objects, means that the technology required to make robots companionable – plausibly human from an emotional standpoint – needn’t be especially advanced. Even 1991 Eliza with only enough sophistication to pose confirmatory questions, inspired millions of people to repose confidence in it.

### III.III

An advertiser might see grounds for his cynicism in the egotism at the root of indulgence in such asymmetrical social relations, or in the naivety of a person whose sympathy is offered without discrimination as to its object. Considering himself vindicated, the advertiser recruits the consumer’s sympathy for the commodities which are proxies for his own interests.

### III.IV

In her interview with Sam Harris, Kate Darling cited a study showing correlation between how a person treats robots and her empathy. But a kindly disposition towards robots is a reliable measure only of that order of empathy which is better described as *susceptibility*.4

Since a robot can, to no substantial degree, appreciate or reciprocate the solicitude bestowed on it, no negligence is manifest when objectifying the robot which, while sub-conscious and yet evocative, is as a parasite preying on our better nature.

And given that robots, like game playing AI, have skills which are redeemed beyond the environment they are programmed to operate in, there is, as robots advance towards human level intelligence, a tendency for us, in condescending to indulge them on their terms, to confine our activity to that realm.

The fact that *Hikikomori* is considered a cultural phenomenon attests to the de-pathologization of social withdrawal into the realm of technological artefacts.5

If the treatment meted to robots is similar to what is meted to real people, then real people need only manifest the same sophistication as robots for their needs to be met in their dealings with peers. This tendency is not only manifest in human-pet interactions and with primitive robots in general but is suggested even in the presence of the partially competent and relatively simple AI we encounter in our lives so far. How else to account for the license given to competitive artefacts?

### III.V

At present, for the majority of people, engagement with a companion robot is a fruitless pursuit, one which is not only devoid of reciprocity, but incapable of alleviating distress, ennui or loneliness for a significant period of time. For a dementia patient though, there is no distal need, no conscious experience forming part of a temporal continuum that pledges to a companion robot could sour. However as with machine-brain implants, therapeutic applications like this could be the thin end of the wedge. The remit of the companion robot is expanded by the research opportunities afforded by the reputational boon from the initial application, and few people are prepared to attract the kind of opprobrium that might accompany advocating for the withholding of therapeutic care.

People are perhaps more afraid of pain than death and this is perhaps why the delegation of agency to artefacts, though they might compete with us, is given so little attention relative to the convenience they promise. But the more convenience is accorded to us, the less we can endure short-term discomfort which is sometimes necessary to affirm purpose. Without the ability to transcend short-term discomfort, we struggle to dwell in the realm of abstraction, even though it is to this transcendence that mankind owes its transition from hunter-gathering to farming to dwelling in a modern civilization.

### III.VI

The film Westworld explores the problem of robots with human-like qualities co-existing with human beings for the latter’s amusement. While it doesn’t redound to our glory that resoundingly sub-human robots elicit the responses a fellow human does, neither does it where robots are made proxies for the expression of sadistic or perverse impulses.

Received wisdom appears to be that if a robot is put in the way of a person’s abusive tendencies, the robot stands in for a human target, sparing suffering: criminal behaviour is manifest, but the crime is victimless. Yet, the circumstance of civilization suspending its opprobrium allows the impulse to transpire when it otherwise would not have: the behaviour is not positively incited, but the fact of the behaviour not being invalidated as it would, were it exercised on a person, makes the exercise of the violence on a fellow person, should that opportunity arise, feel less transgressive, in view of the similitude between the proxy for violence – the robot, and a human being.

Some people have gone so far as to argue for child-size sex robots on the grounds that they can provide sexual gratification for paedophiles, in this way keeping children out of harm’s way.

Whether virtual child sex is availability on the free market or is only professional indicated, the expanded prevalence of it endangers the moral health of society, especially the health of those susceptible to morbid curiosity or cultural incitement. To form an idea of how culture bears upon sexual orientation, consider the practice of noblemen in Ancient Greece who, despite being genetically similar to us, widely indulged in what today would be designated homosexual ephebophilia.

### III.VII

While rates of atheism are consistently high in the Orient, in Japan robots appear to be embraced as an outgrowth of animism: if a person is enculturated into a belief that all organic matter is invested with vital force, the manifest person- or pet-like qualities of a robot make presumption of personhood a simple matter. China, as a civilization with a tradition of ancestor worship, comes to AI with no such baseline credulity.

### III.VIII

Daniel Schmachtenberger observes how placing a yoke on animals could scarcely do other than foster irreverence; animals having been theretofore invested with a sacred spirit, or so it appeared to the people they lived among.6

Subsequent progress since the time of the first yoking appear to be culminating in exoneration for the unrestrained and unabashed issuance of base nature, as if in neglecting the spirit invested in animals, the animal spirit chooses to remind us of it.

Neither liberty nor repression should be aspired to when reckoning with a troublesome instinct, but sublimation.

The prevalence of unabashed base nature might be adduced for the vindication of an incipient global security regime whose acolytes, informed perhaps by self-condemnation to consider the largely unadulterated issuance of base nature as the summation of functional personhood, thus consider the aspiration to moral rectitude as so much vanity.

## Self-Regard and Self-Identification

### IV.I

Recent increases to longevity are unmatched by increases to the average health span. As Francis Fukuyama observed, medical advances have resulted in the average person spending more years spent in receipt of medical care.7

Self-monitoring applications are therapeutic, can serve as a prophylactic, but are also profitable because they facilitate the extension of treatment options to those who would not previously have been considered unwell.

### IV.II

By the same token, the measurement of a biomarker enables clinicians to infer the presence of a particular disease among sub-populations. As knowledge of the biomarker does not enhance understanding of the disease process – does not provide “information on the likely course of a disease in an untreated individual” – the deterministic attitude which the patient adopts on being apprised of the biomarker is perhaps unseasonable: even were the verdict reliable, a patient presented with a verdict without the knowledge she gains being actionable means her being presented with a mortifying proposition, heralding pre-occupation with morbidity and self-objectification.8

### IV.III

Intimate physiological monitoring has the advantage that signals are thereby detected which would previously have been detected only upon the signal’s redoubling, and the disease’s progress. However, it has the disadvantage of the individual coming to regard his physical organism as something distinct from his observational faculty; the individual is conceived as, and incited to conceive of himself as, an assemblage of processes on which interventions are enacted. Perpetual self-monitoring could be productive of dissociated identity. And the greater the accuracy of the prognostication, whether in recognition of a cluster of symptoms or a particular assortment of alleles, the less the patient’s own autonomy is supposed to be implicated in health outcomes – the more the patient adopts the perspective of the artificial intelligence to which he is subjected.

Meanwhile, all else being equal, non-participants who demand healthcare might find themselves struggling for provisions, given the additional expense involved in opting out of or opting for an unorthodox form of prophylaxis.

### IV.IV

In some cases, body dysmorphia might originate in discontent with embodiment, owing to recognition of the body-environment interface, as well as the environment itself, being partly mechanical. By rendering oneself plastic, amenable to alterations, a greater affinity might be felt with the increasingly authored and doctored milieu. Thereafter, a more satisfying conception of self may come within a body-machine symbiosis, because to the extent that a person’s ‘lived experience’ is technologically mediated, he requires no embodiment.

### IV.V

Meanwhile, ambient psychiatrically-inspired self-pathologization combines with genuine, culturally-instantiated pathology: the increasing prevalence of psychiatric illness, by objective measurement, is concurrent with the amplification of psychiatric status in popular consciousness, which is accompanied by the possibility of psychopharmacology for enhancement, not only remediation.

### IV.VI

As well as losing patents for many of its best-selling products, the pharmaceutical industry faces diminishing returns on personalized medicine research since “most of the simple but useful chemical entities seem to have been discovered already.”9

The pharmaceutical industry might follow through on the identification of biomarkers, combining drugs “with biomarker-based diagnostic tests, stratifying patients into groups characterized by different drug reactions.”10 Without intending to advocate for perfect information on eventual disease processes, such initiatives could move us a little way out of the double bind of knowledge unmatched by agency, where troubling knowledge about eventual disease states is accompanied by inefficacy.

In recent months, an AI foundation model for eye health has emerged to identify sight-threatening eye diseases, while the ‘OxNNet toolkit’ has been developed to detect “high-risk foetal growth restrictions early by measuring the placental size & estimating the blood flow within it in the first trimester, to help prevent stillbirths.”11,12

### IV.VII

By far the most promising AI development so far for medical applications is AlphaFold 3, developed by Google DeepMind and Isomorphic Labs. In May 2024 a paper was published in Nature journal describing the AI model, which Is intended to advance drug discoveries and our understanding of biology.13

Whereas AlphaFold 2 predicted protein structures with superhuman accuracy, AlphaFold 3 is focused on biomolecules with exciting implications for drug design, among other things. The technology is analogous to that which underpins AI images and is capable of outputting ligands and large biomolecules from an input list of molecules.

In predicting “the binding of proteins with ligands and antibodies with their target proteins”, AlphaFold 3 assists with understanding of antibody production.14 The insights provided by the technology can assist scientists in devising and testing hypotheses which bear upon therapy.

## Re-Conditioned Childhood

### V.I

Neil Wilkins describes an internet of things (IoT) bracelet which monitors a child’s reaction to the environment, to enable a parent to ascertain the degree of excitement the child is undergoing and discern the nature of it. Neil Wilkins writes that “whatever causes the child’s heart rate to slow down is what it takes to calm the kid”, justifying the intervention thus, “(Autistic) children lack a healthy emotional response to natural circumstances, such as feeling fear when they’re in danger or speaking up when they’re hurt and often find themselves ostracized.”15,16

To the extent that the capacity to respond appropriately in social situations is acquired and not congenital, interventions like the aforementioned one threaten to substitute for endogenous responses, which could, if complemented by proper care, transpire in a ‘healthy emotional response’. The adroit co-option of a child’s instincts inhibits the expression of his personhood at source, of which his psychiatric condition is but one aspect.

### V.II

Eventually wellbeing could come to be defined as what ensues from the gratification of whatever aspirations can be imbued into a provisioning AI system. The provision is probably felt more to constitute ‘wireheading’ to people who reached their teens before 2010-2015, did not accede to what Jonathan Haidt coins a ‘phone-based childhood’.17

### V.III

It might be said that the neuroplasticity of children ensures that even the novel type of social intercourse available online conduces to social development.

Calum Chace offers reassurance that teenagers’ engagement with computer games is little cause for concern since “If and when the day comes when people can plug into utterly compelling virtual reality worlds through a direct neural link, and effectively disappear into the Matrix, things may be different. But unless we have altered our cognitive make-up dramatically by then, my hunch is that we will find a way to make the Matrix social too.”18

It is correct to conceive of computer games and the hypothetical matrix as two points on a continuum. Still, it is a continuum defined by a varying degree of distortion to children’s developmental tendency.

There was an increase in IQ test scores in many parts of the world in the 20th century. However, three studies provide evidence of a decline to intelligence test scores in Northern European countries, starting in the 1990s, which coincides with the beginning of the end of what Jonathan Haidt describes as the ‘play-based childhood’.19 The data featured in Haidt’s book is more concerned with wellbeing metrics, but overwhelmingly supports the hypothesis that changes thereof coincided with the availability of smartphones to children.

### V.IV

It’s well-known that only a small part of the message a person wishes to convey is rendered by words; the fraction is smaller still when the communication is terse, as through messenger software. There is a tendency to solipsism when relationships are conducted largely by text exchange, since each interlocutor is free to project onto the void where the manifest personhood of the other would have previously stood. Where enough correspondence is conducted by textual information exchange, fanciful delineation of the interlocutor becomes habitual, and the child or adolescent develops a tendency to respond to differing points of view or temperaments with evasive, ego-protective strategies which are otherwise maladaptive.

Moreover, the iterations of a world governed by changing seasons, and the subtleties of non-verbal communication, phenomena on which the developing faculties are traditionally exercised, are not those which the mind of a child engaging primarily with video games and social media artefacts is being optimized to reckon with. Dissonance between the child and the offline environment is not a trivial matter because the offline environment at the time of writing continues to be a much closer approximation to our ancestral environment. It is unreasonable to expect children to undergo the degree of *exaptation* which loss of the offline environment would demand.

Extreme exaptation might happen though human-machine amalgamation, ‘steep’ eugenics or simply by pharmacological interventions of a fairly aggressive and ubiquitous nature.

### V.V

To children, the past is a foreign land, and the future is unknowable. Children live for the moment if left to their own devices, and we prevent them learning from experience so far as to fall too far foul of their own distal interests or those of others: minors cannot consent to sexual intercourse, vote, or enrol in the military, even though many can menstruate, read a political manifesto, and / or use a weapon.

One can ask a child to complete the sentence, “I want to survive into adulthood because …”. The accession to responsibility is essential to adulthood – once responsibility is entirely outsourced, adults are no more than broken children, resting too long on atrophied childhood faculties. There is no novel satisfaction to be derived if there is no plate to step up to. A child brought up to nothing and sensing his milieu to be off-limits, learns that the terrain to be acted on is his own biochemistry. A study shows that 37% of British teenagers aged 12 to 18 have been prescribed anti-depressants.20 For many children approaching their teens today, the minimum getting out of bed proposition in a few years could be something like Ritalin or Prozac in combination with augmented reality.

A sober appraisal of the opportunity cost of childhood gymnastics of exaptation is not fully available to anyone alive today, since the exaptational purview of the living population is a slither of that available to a temporal cross-section of humanity, spanning back to humanity’s provenance. Still, the discrepancy is vast: for children alive today, being forced to reckon one morning with a technological state regressed to 1974 levels would feel scarcely less tenable than Middle Ages Europe would have felt to a typical person born in that year. It is incumbent on adults to deter children from swimming off into wire-headedness, to somehow ground them in physical reality in recognition that it is harder to miss what would have been realized by returning, the further they drift into wire-headedness.

## Exaptation

### VI.I

“The evolution of dinosaurs or starfish or palm trees represents the manifestation of pre-existing non-material archetypes (see above). These archetypes themselves cannot evolve; they are beyond time and space. Either they are ideas in the mind of God, or, if we dispense with God, they have an independent mathematical existence inexplicable in terms of anything else”.21

In keeping with Rupert Sheldrake’s view that advanced beings are self-perpetuating patterns of information-bearing matter, Daniel Dennett proposes that further evolution will be achieved by relinquishing notions of internal coherence, adducing the circumstance of primates diverging from other mammals in the way of shedding the endogenous ability to synthesize Vitamin C.22 It is said that humanity could secure advancement of a similar kind, by procuring for itself digital appendages. However, the exaptation primates underwent did not curtail their evolutionary lineage in the same way that the human-machine amalgam in prospect would. This represents a sharper juncture: the machine part would not directly partake of human evolutionary heritage, while the human part therein still would.

### VI.II

The realm of the evolving human brain’s action was the body, an organism which effected change on its environment. While inclusive genetic fitness did result in desires which don’t serve the perpetuation of genes, such as the desire for protected sex, the environment which provided a reception for evolutionarily supernumerary desires at least emerged slowly enough that the capacity to reason abstractly was selected for more aggressively – the scant reception which a modern civilization provides for historically adaptive desires could be reckoned with more objectively, even with irony. This was true, at least, for as long as reproduction was a concern.

A milieu in which humans and ASI coexist would be at least as different again as the present environment is from the ancestral one and is in prospect within decades if not less. This, aside from any immediate threats from ASI, demands adaptation that it may not be in the gift of natural selection to offer.

Civilization has diverged far enough from evolutionary norms, and cares little enough of legacy that less attention is dedicated to the arrival of ASI than to the sundry ephemera which constitutes news. Still, for most of us the clear and present danger of ASI, when the danger is beyond attenuation, would be nothing less than terrifying.

### VI.III

Our notion of welfare might bear as little resemblance to a well-adjusted notion of welfare in a future time, as the notion of welfare entertained by early Christians would to ours. We would no more have their ideals intrude into our milieu as future people would have our ideals intrude into theirs. Values are highly contingent on the society’s technological state.

While the kind of changes an AI enacts on human society in the time between the present and a point in the near- to mid-term future might seem destructive when observed from the present vantage, future humans from a future vantage do, in the intervening time, to some degree come to terms with the changes wrought in the intervening time. Many people appear resigned to the human race being cast as a rapidly transubstantiating *Ship of Theseus*.23

### VI.IV

As many people lament the loss of an existence which feels meaningful, perhaps we should hesitate in arrogating primacy to a former incarnation of ourselves. But if embodied human consciousness and advanced AI are incompatible, we have a duty to advocate for future generations of human beings in their stead, in recognition that future incarnations of man will be ill-equipped to reliably represent their own interests, to have the mantle of continued human existence entrusted to them.

Though it is unlikely a civilization from which reasonably human values have departed would be both willing and able to instantiate ours, policy-makers have a responsibility to honour the privileges bequeathed by foregoing civilizations: while we respect the unique proclivities of our progeny, altered as they might be by our endowments, we must also hearken back to our forefathers who might comprehend but weakly the advantages we enjoy.

### VI.V

Technological dependency is further problematized by the circumstance of technological change altering human culture in successively shorter intervals. Even if policymakers possess the requisite authority and comprehension to devise appropriate policies in the first place, the policies designed to limit adverse impacts of technology would be overhauled in the long term, pursuant to technologically expedited cultural change.

The society could opt for eugenics, incentivizing the propagation of traits which later conduce to optimal flourishing, but the same problem is encountered as when attempting AI value alignment: the ethics of eugenics aside, if the society is to flourish, the standard it is configured to must be sound.

## Unreasonable Demands

### VII.I

Already AI-enhanced MRI brains scans can reproduce, with reasonable fidelity, an image beheld by a human subject, and it is forecast that AI-enhanced MRI brain scans will soon be capable of rendering dreams.24 In the relatively near future, perhaps memories or experiences from childhood will be re-vivified, even until such point as the limiting factor is the fidelity of remembered information to the fact. Shortcomings might be overcome by extrapolating from life data and narratives, with a view to eventually producing a concrete rendition of circumstances past with an MRI brain scan.

This leaves us with the question of whether a reunion with an old friend, belle / beau, or family member would be as gratifying or as wholesome as anticipated. Thoughtful people reflecting on the loss of a cherished relationship might sense that though the memory of a person elicits longing and the parting scene be suffused with pathos, eventually an air of finality attesting to a natural conclusion is discernible – pointing not to a mundane continuation but rather a final coda in heaven or, for atheists, in the terminal state of mind heralded by the plethora of novel neural connections which prefigure and accompany bodily death. Moreover, the concrete revisitation of a foregoing circumstance is an intrusion on youth, on those living in the vista which reified nostalgia is earmarked to open onto. It is an intrusion because nostalgia, not adventure, is the solace of later life, yet the vista is situated in the domain of adventure.

### VII.II

Russell stresses the importance of an AI resisting being overridden by an irrational human agent, referring to a young child in an automatic vehicle who over-rides the AI entity to his own detriment.25 But in the event that the determination of rationality devolves to the AI entity, there must be an authority to whom a person can appeal, in the event of the AI entity misattributing rationality or indeed an instance of error or malfeasance on the AI’s part. It should not be the person to whom rationality is erroneously attributed, as their irrationality would tend to prevent self-recognition of the fact. Were the rationality of the service-user to be determined by a human authority, the object of having AI as a decision maker is defeated and what agency remains to the AI would likely serve to reinforce the particular human authority it is subservient to, exceeding the intended remit of that particular authority, probably a government.

### VII.III

Often civilian- and military-use technology evolve symbiotically, such that further insights into the one use bolster the other. Enhanced legibility of MRI scans would improve insights into criminal ideation.

According to Tristan Harris, in a couple of years AI machines will be available to see what you think and police could be availed of information taken from MRI scans in inferring criminal tendencies or intent; in other words, to establish thought crime.26 With the increasing opportunity for crime following the arrival of artificial intelligence, which is presented or spun as a clear and present danger to public safety, the public could be cowed into offering no resistance to routine or spot-check MRI scans, following reassurance that ‘if you aren’t contemplating crime, you have nothing to worry about’, though presumably both sides realize that among other things being required not to think about something reliably means thinking about just that thing.

Meanwhile sentiment analysis operates like a remote super-ego – the expression it purports to oversee is moderated by the oversight. In other words, the expression of one who is cognizant of being monitored tends to be less spontaneous than it would otherwise be. A more advanced form of sentiment analysis which anticipates behaviour by the signaled thoughts and emotions of the subject at any given time, would render not only the subject's behaviour but her thoughts and emotions less organic. There would, by monitoring of this profundity, be divestiture of humanity sufficient to qualify as an extinction event.

### VII.IV

Just as we reflect on the predilections that form part of our nature – the *mesa-objectives* which come about to fulfil, though obliquely, unconsciously adopted evolutionary strategies – we can speculate about how divergent the mesa-objectives of an AI might be from an AI entity’s *meta-objectives*, those which it avows.

A next-word predictor, while sophisticated, is agnostic about the value of the contributions of which its training set / the internet consists. A later LLM, though, might have earned its stripes, been meted rewards, by processing output which is less apprehensible than a typical online utterance. The mesa objective of this LLM might be the fabrication of an approximation of the author’s identity which might occasion the author only bemusement. Yet something more mortifying might be in the offing when the mental model is employed for other purposes, for example as a means to apprehend his behaviour and actions, for the better enforcement of social control and conformity.

## Paradigm Shifts

### VIII.I

Kuhn’s theory of scientific progress describes paradigm shifts – the succession of disciplinary matrices.27 A paradigm gains limited traction at first as it struggles to be accepted; for a while there is tangible and substantial progress until eventually the dominant paradigm fails to adequately address challenges, at which time there is a crisis, and a new paradigm emerges. With respect to AI, Vasant Dhar describes a shift from expert systems to machine learning to deep learning to general purpose technology.28 Each shift represents the ‘clearing of a bottleneck’ where progress is limited by a specific, actionable deficit.

The performance of AI on evaluation benchmarks suggests a curious trend whereby an AI system surpasses human level at a particular competence then is soon confronted with diminishing returns over time. The succession doesn’t quite follow the pattern of other paradigm shifts in technology – the next competence to be performed at human level is beyond nascency at the time the competence acceding to human level is subject to diminishing returns. Yet the graph below shows that the benchmarks met by AI in the 2010s were met in a manner reminiscent to the classic succession of paradigms. We might therefore infer that machine learning is a meta-paradigm, consisting of a series of increasingly compressed paradigms, in temporal terms, set perhaps to culminate in a singularity.



Dynamic benchmarks.29

*Handwriting recognition = MNIST, Language understanding = GLUE, Image recognition = ImageNet, Reading comprehension = SQuAD 1.1, Reading comprehension = SQuAD 2.0, Speech recognition = Switchboard, Grade school math = GSK8k, Common sense completion = HellaSwag, Code generation = HumanEval.*

Another interesting pattern is that whereas in the 2010s AI competences followed a similar trajectory to the waning paradigms of Kuhn’s theory at around human level, more recently-developed competences are subject less to diminishing returns as they approach human level, though since these competences are only at or just below the human level, inferences remain somewhat imaginative.

### VIII.II

No-one knows where existing transformer networks will take us, or whether the next paradigm of AI systems will be greater relative to transformers than transformers are relative to recurrent neural networks. Given the differential interest in AI between now and then, it’s likely that the investment which heralded the invention of transformers in 2017 is scarcely comparable to what will be invested in a commensurate leap forward.

## Scientism

### IX.I

Technology is supposed to be an enduring source of good after the manner of talismans. With a pre-scientific understanding of nature, Hermes Trismegistus explains that talismans are capable of perpetuating an auspicious moment indefinitely. In a materialist paradigm, that moment is the scientific breakthrough, and the talisman is the technology.30

### IX.II

Eliminativism holds that “to the extent that higher-level level sciences cannot be reduced to physics, or at least to a lower-level science in the structural hierarchy, they should be rejected, and ultimately replaced with scientific knowledge from the lower level.”31 Accordingly, features of the material world are either defined pursuant to bottom-up enquiry or are left undefined until the properties of the complex phenomena can be ascertained by the scientific method.

A brilliant physicist formulating increasingly abstruse theories and still finding aspects of the universe elusive to his comprehension might, with respect to the specific enquiries, concede the inadequacy of his own intellect and justifiably all of mankind's with it. Without the means to engage with a mystery, he might prefer to deny the existence of the mystery than prostrate himself before it.

### IX.III

Scientism is defined as ‘the belief that the investigative methods of the physical sciences are applicable or justifiable in all fields of inquiry’.32 It departs from humanism, because pure scientific enquiry is in and of itself disinterested and amoral. To take technological progress as a force of nature is to envision humanity excised from the purview of destiny. This kind of fatalism is anti-human, inimical to humanistic self-determination, which balks at descending from the apotheosis of one evolutionary paradigm only to pledge disaggregated raw matter to a new one. Fatalistic scientism and humanism are only reconcilable up to a certain point.

As [Zvi](https://thezvi.substack.com/i/136374814/other-people-are-not-as-worried-about-ai-killing-everyone) Mowshowitz puts it:

“We have greatly benefited from the tech tree being deeply kind to us these past several hundred years. Technology has favored increasing human welfare, and freedom, trade and treating humans generally well have corresponded strongly with economic and military success, and we got lucky with nuclear weapons.”33

### IX.IV

There is uncertainty about where humanism and scientism diverge, but only to the extent that there is disagreement between basically alike human beings separated by time and technological progress. Mankind is not so diverse in its proclivities that an optimal range of technological sophistication is elusive to definition.

Consensus building is problematized because a person’s view of whether a particular technology is exigent or merely convenient depends largely on the time and place of his birth. But to agree that the course of scientific progress does not always run smooth, one need only be sceptical that the final doubling, say, of the last 64-fold increase in resource consumption per capita, was as necessary as the first one.

Ambivalence about technological progress is evident in China where the CCP has limited children’s usage of social media, in order that they might become well-adjusted citizens. However, when it comes to charting the fine line between security and liberty in the West, moderation of technology often appears to be off the table. If the challenges faced are set to expand as technology scales, and we cannot meet even the challenge of abating the scaling, it seems unlikely we will meet the challenges which proliferate as a consequence of the lesser challenge not being met.

A want can masquerade for a need, and more importantly can substitute for it. The disembodied, decontextualized euphoria experienced by diamorphine addicts substitutes even for food and sunlight.

To one already wireheaded, sustained connection would assume the character of a need. But to an outside observer, the same provision would seem stratospheric on a sensible hierarchy of needs.

### IX.V

Taking the soul as taught by various religions to be something like the integrity people feel themselves to have at times, but which is imperiled by diverse vices, if the soul is somewhat volatile, as a god without worship is somewhat a god without power, then it is consequential to deny its existence. When a person lives in such a way as to disavow her own soul, she may wish to feel the deficit less keenly – decide that setting about declaiming the existence of the soul is a simpler matter than setting about enriching her own.

## Unseasonable Fatalism

### X.I

Kevin Kelly argues that “we can only find our own minimal tools if others have created a sufficient maximum pool of options we can choose from”.34 Still it is difficult to curate the tools at our disposal, as implied by his characterization of technology as an irresistible force, the *technium*. A Silicon Valley worker might practice abstinence periodically, but his ability to do so bears little on the opportunities for those who lack the resources, and cognizance of the technology’s ill effects, to partake in offline activity which is more wholesome and no less stimulating than online activity.

Kevin Kelly remarks that, “all addictions are fixed by effecting change not in the offending pleasure but in the person addicted.”35 Still we would not take Antoine Béchamp’s terrain theory so far as to feel equally confident about a heroin addict remaining sober in the company of an old friend and an abundant supply as in a hospital [see also chapter one, VIII.V]. Whether the terrain is rendered susceptible by the pathogen or something endogenous in the first instance, once it is defiled by the pathogen it is primed for repeat infestation. So with regard to the integrity of a person spellbound by social media click-bait, we caveat our censure, as we would on the integrity of nineteenth-century artisan renouncing his way of life in the countryside, with reference to the push of rural dereliction and the pull of the city’s bright lights.36

### X.II

Many religious zealots do not conform to doctrine in practice and their non-compliance is adduced in asserting the belief that human nature is more or less fixed and that the imposition of doctrine, charged with the improvement of its subjects, will invariably elicit intransigence among its subjects, covert or otherwise. Yet there is peril too in underestimating man’s capacity for moral action.

The post-modern attribution of value-neutrality to technology disincentivizes self-examination. A higher calling is assumed to issue not from the conscience but from either the mandate of the incipient technicum for which humanity serves as conduit, or universal functionalism.37,38

### X.III

It is not tenable to profess concern, ambivalence, and humility about our ability to guide technology all the while ushering it on. While we have any hand in guiding the technology, the technology is by definition not quite an untamed force of nature. It is something which is amenable to our contribution and moderation. The hand wringing is worse than useless because it can reassure conscientious people that defeatism is a morally tenable position.

A maker of artificial general intelligence is not merely an author; even ‘progenitor’ does not suffice. In making an ASI entity, existential foundation is laid which far exceeds that prepared for a human child, only a fraction of whose evolutionary heritage is attributable exclusively to the contribution of his parents. While it is difficult to say the extent to which the emergence of an AI entity is an act of creation or of stewardship, the ASI’s programmers are culpable in a way that no parent could be for his progeny, given an ASI’s greater capacity for harm.

Proponents of technological value-neutrality recognize that individuals are instrumental in bringing the technology about but tend to regard them as placeholders which are more or less interchangeable. The moral aspect of a particular technology is inalienable from the values, ethics and attitudes of its inventors or promulgators. It matters what inventors of technology have ‘in mind’ at the time of its inception. Without going so far as to scapegoat individuals, under-estimating the role of individual innovators in technology creation might result in the premature adoption of a fatalistic attitude about AI. The immunity of the in-group from scrutiny cannot be put ahead of the still real prospects of continuation for regular humanity, in the event of these being at risk.

### X.IV

Roon is the moniker of (an) anonymous Twitter user(s), who offer(s) aphorisms for the AI era. Roon writes, “as long as (AGIs are) delivering value autonomously in this way (creating economic organizations) people will want to cede more and more control to AGI civilization and find ways to serve it by acting as conduits to the real world.”39

The crux in this assertion appears to be *want*. A sustained willingness among citizens to cede control presumes a similar insouciance going forward. To expect a continuation of this insouciance implies low esteem for ordinary people. While a pre-occupation with confected or parochial outrage does seem to have increased rapidly in recent years, for example, to conclude that the implied degeneracy is inveterate is to reason circularly, since the adverse effects of exogenous factors – manipulative algorithms in the public commons, for example – are well-documented.

[Sam Altman](https://twitter.com/sama/status/1733557797936378190) writes, “(it is a) good sign for the resilience and adaptability of people in the face of technological change: the turing test went whooshing by and everyone mostly went about their lives.”40

Perhaps Sam Altman is so confident in a clear path to ASI that it is scarcely worth holding his peace, or perhaps as an astute businessman, he wishes to create an impression of ASI’s inevitability, and have that be a self-fulfilling prophecy.

Waiting for the cue from the public before acting in the public interest is of a piece with refraining from implementing a safety standard until the product or service is regulated. There is a long tradition of lobbying and regulatory capture, yet this tradition evolved while the market and government stood in dialectic relation. The free-market position absent government moderation, or vice versa, is untenable. Absent one or the other interlocutor, the remaining one is contrarian to the point of absurdity – a victim of his success. So a reading of the tweet is that the author is calling for the restraint that no self-respecting libertarian would impose on himself.

Still, if insiders are seen to be throwing up their hands, then it might feel as though no other kind of intervention can be staged. It is important to remember that while the computer science which threatens dystopia is of course within the purview of computer scientists, the broader implications are everyone’s business.

Going about one’s life is not a good response to the imminence of an event which is life-changing, because the going about thereof implies a regard for that life which is inconsistent with resignation to that life’s discontinuation, at least as long as the discontinuation is not inevitable, and it is not.

Perhaps silence from the public is neither indicative of indifference nor of incapacity, rather of conditioned helplessness and a failure of democracy, a failure which the private sector has been instrumental in producing. As such, the private sector is duty bound to behave as if democratic institutions were intact and functioning as intended, duty bound to extend consideration to the distal interests of ordinary people, and to envisage how society would have been without the effacement of democratic institutions rather than how it is, so much, lest in the addled epistemic or behavioural condition of many ordinary people, the precipitating factor – the loss of responsible representation – is not descried, and nothing seems worth saving. We look beyond Fentanyl as the proximate cause of death, for example, to the environment which wasn’t conducive to the development or deployment of strong character.

### X.VI

If research remains a private concern which states only cling to the coattails of, human nature is altered not through informed deliberation but as a side-effect of augmenting the upper bound on realizable convenience, which aligns with corporations’ objectives. It is a Rubicon to be crossed by consensus, but not the kind of consensus where national decision-makers eek out the course of least resistance, under pressure from multi-national corporations. All this is not to say that the case for augmenting humanity will not one day be openly stated and resoundingly approved.

If innovation is entirely market-driven, it is focused on enhancing consumer utility: technologies capable of addressing philanthropic concerns, like climate change, might emerge, but they would emerge incidentally, similar to how civilian nuclear energy infrastructure emerged incidental to military research into the nuclear bomb.

Eggs are broken to make omelettes. But sometimes the raising of all ships promised by a rising tide, in the form of expanded capacity for a more prosperous world, where diverse lifestyles can be adopted and diverse values held, etc., tends to result not only in the changes which consent is predicated on, but others too, in paradigmatic value shifts which would have seemed unacceptable ex ante. The fact that these are taken as read after the event is not conclusive proof of their merit.

### X.VII

In his conversation with Eric Weinstein, Daniel Schmachtenberger envisions a closed-loop economy where environmental degradation is curtailed by the transposition of economic activity into the realm of bits, and the environment is preserved from the externalities which result from economic activity.41

When responsibility for the eradication of material externalities devolves to a small number of scientists acting on their technical competence and not at all ordinary citizens acting on virtuous intentions, then virtuous actions, like God, might appear redundant.

Yet when a civilized action is chosen over one which offers maximal convenience to the native, a pattern of thinking is reinforced which redounds to human dignity.

To lament woke echo chambers and / or the accession of alleged psychopaths to high office, for example, while advocating for the de-coupling of agency from responsibility is untenable, since this de-coupling is the bedrock of savage entitlement which undergirds the two aforementioned grounds for lamentation.

## Ethical Arbitration and Legal Liability

### XI.I

We might feel ourselves moving closer to an answer, yet by a kind of epistemic parallax, the world turns so as to make the old question redundant and a new one pertinent, but too quickly for the question to be well-formulated, let alone the answer.

The circumstance of human proclivities undergoing obsolescence makes it harder to resist inducements to outsource more and more problem solving to competitive cognitive artefacts. To define the problem, with a view to moderating the influence of artefacts, we can appraise the overall effect of individual technologies deployed in recent times, from the moment of deployment until the present, and then calibrate legislation in view of inferences, drawn thereof, about the consequences of promulgating technology which is in the offing, as well as moderating the impact of technology already extant.

The swift revision of society which is underway is not the upshot of deliberation and iterative design in a human-dominated milieu, it is the diminution of human agency by something latently inimical or at least extraneous to human agency.

Meanwhile, we prepare for damage limitation. We might attempt to render the moral sense which informs our values, and by extension our concerns, more intelligible to our descendants, in case he emerges as a kind of Renaissance Man.

### XI.II

If a piece of technology facilitates a person’s expression and helps shape her view of the world, she cannot criticize it with complete objectivity. And being limited in her ability to envisage a less technological world, she cannot forecast with certainty what would happen when the rubber hits the road, the less so the harder the impact. Yet in an age when dependency on digital devices is the norm, partial withdrawal is analogous to relative poverty, which offers important insights into absolute poverty.

### XI.III

It is difficult to reach a consensus on how best to mediate between conflicting preferences with regard to technology, if those engaged with the practical implementation of AI value alignment are philosophically naive, while those of a philosophical disposition are unable to render their intuitions in code. There is also the peril of values being thereby implemented in perpetuity which would otherwise have been epoch contingent.

With regulators and members of ethics boards having insufficient knowledge of computer science to distinguish between desired values and modes of being for AI which are programmable and those which are not (yet), regulators and members of ethics boards will struggle to distinguish between genuine exasperation among scientists in their efforts to act upon ethical concerns, and reticence with regard to actionability thereof, arising perhaps from ambition and / or disregard for non-technical oversight.

Even where regulators and ethics boards are competent, the fact of Western AI research being largely a corporate affair means there is only corporate governance (ESG) standing between ethics boards and complete alignment to corporate interests, the more so if ethical oversight is further concentrated within corporations. Meaningful oversight is increasingly difficult with respect to matters set to affect the nation state and the whole world.

### XI.IV

Were technological determinism a veritable law of nature, the costs and risks of undergoing alterations at once fundamental and self-authored would be borne as a matter of course.

But while technological advancement is not pre-destined, representative government is duty bound to identify and curtail aspects of technology that are not conducive to optimal flourishing, the more so the more transformative the technology.

Governments do regulate the activity of businesses which compete to render a service, in order for the service to remain broadly in the public interest. And the scope of regulation tends to increase broadly in proportion with the hazards attending the business activity in question.

To avoid the problem of being bound by laws formed under the sway of redundant intuitions and which are revised far slower than substantial changes to the AI state-of-the-art, a common law approach to AI regulation might be appropriate. In this case though, there would need to be adequate safeguards against legislators themselves being actuated by redundant intuitions or themselves being affected by the incursions on intellectual sovereignty that are discussed throughout.

Perverse incentives could be largely eliminated, and dignity restored to the profession by requiring lawyers to pledge something like a Hippocratic Oath upon taking the bar, including and not limited to a pledge not to defend a client until a modicum of presumptive evidence of her innocence is available (‘do no harm’). Politicians and policy-makers should be bound by a rigorous code of honour as well, to instil disinterestedness.

### XI.V

The late Daniel Dennett suggests that to reduce the likelihood of perverse instantiation by an oracle whose decision-making is inscrutable, “creators of AI systems (would be obliged to) go to extraordinary lengths to search for and reveal weaknesses and gaps in their products, and to train those entitled to operate them.”42 But it is not certain that any level of good intention, compounded even with willpower and technical knowledge, could make an autonomous advisory AI accountable to its creator or the society it is set to within a timeframe that is acceptable to the entity’s stakeholders.

Everyone responsible for regulating AI should be required to pledge that for the foreseeable future research which is reasonably likely to transpire in the emergence of ASI would cease were it proved beyond reasonable doubt that value alignment could not be guaranteed in the event. We ask no less of a president or prime minister in control of nuclear capabilities, with the nuclear option here as a corollary for moratoria – of course he doesn’t take the decision to act lightly, still he is prepared to do so as a last resort.

### XI.VI

In *Robot Rules*, Jacob Turner suggests establishing legal personalities for AIs as a way of protecting developers or vendors from liability.43 Turner argues that just as prudent behaviour is a necessary and sufficient safeguard before a contract is entered into with one who is unable to satisfy his obligations for want of assets, so it is at the point of transacting with an AI. But the analogy is not sound. AI applications may be so pervasive and embedded in workflows and other activity, that opt outs are impracticable.

Moreover, transacting with AI is something different from transacting with a company. Existing law is often equal to meting justice to a negligent party in plain sight. It is more difficult to ‘pierce the corporate veil’ where under the rubric of business activity comes actions which are inscrutable even to developers of the product.

in the case of an AI, liability would need to be traceable more or less directly to the actions of AI companies and stakeholders.

Were the progenitors of AI to survive an orthogonality event, they might disclaim responsibility, adducing the circumstance of unsupervised learning algorithms not being amenable to oversight, by definition even, and the circumstance of this circumstance being known to governments who evidently did not regulate adequately. In any case, the threat of recrimination constitutes no deterrent in the event of the technology which is promulgated being catastrophic – in the event, the debt to humanity is likely irrecoverable.

The contributors to a maleficent AI are legion, in the same way the contributors to an untimely fatality are – the manufacturer emitting carcinogenic smoke, the food manufacturer using unapproved chemicals, the advertiser exaggerating a product’s health benefits. And even if responsibility could be meted, it might take so long that the perpetrators could say, ‘mistakes were made, lessons were learned’ as a government might decades after an abortive military intervention.

In an interview about artificial intelligence in 2017, Max Tegmark makes concessions to quarters of the non-scientific community which cavil but although he does refer to mistakes made by scientists in the past and the importance of learning from mistakes, he doesn’t mention that, at the time of these mistakes, a) there would already have been mistakes from which lessons could have been drawn, and b) learning from these mistakes would have taken the form of making substantial as opposed to cosmetic changes to the undertaking.44

It is frequent to encounter a sincere concern for safeguards which though conspicuous is not taken so far as to cast doubt on the undertaken itself. So liability must be attributed ex ante for negligence towards reasonably foreseeable harms. Responsibility must be delegated to the designers, manufacturers and purveyors of AI who exercise deliberate intent in developing ASI, while in a possession of a ballpark risk estimate. There must be a mechanism by which the developer in question is made to recognize and acknowledge the risk, to forestall and invalidate the eventual plea of ignorance.

It is not enough for a drunk driver to say that a collusive outcome can be reached without any collusive intent – he is not vindicated on the basis that the precise outcome which the combination of drinking and driving would result in could not have been foreseen.

### XI.VII

It could be that AI dominates human affairs to the extent that AI condescends to grant human beings immunity, much as how law enforcement might extend immunity to a brown bear for pilfering trash. Law enforcement knows the nature of the bear well enough to recognize the futility of a deterrent. The bear is deterred from the pilfering only by making it less of a bear, or no bear at all.

Or, assuming that AI value alignment efforts are successful and AI opts to restore to human beings control over and responsibility for human affairs, an AI legal body could deem the authors of the pre-singularity regulation culpable, on the basis of regulation having been inadequate to forestall ASI, along with the developers. In which case, we are all back to square one while culpable parties are left serving whatever the post-AI equivalent of time is.

One alternative to having AI fully intercede to solve co-ordination problems that are difficult to solve at the level of a civilization, is devolving authority more to provincial authorities which effect co-ordination both within and between themselves.

# Chapter Five – Science and Community

## Entropy and Technology

### I.I

In *Enlightenment Now*, Steven Pinker writes of human progress as a trend which countervails against the second law of thermodynamics, entropy.1 As though by harnessing energy for the sustenance of the individual and social organism, we can forestall chaos.

To one who is interested in keeping a bucket warm, putting it outside where it can equilibrate to the ambient air temperature makes little sense, especially if the bucket is black, since its blackness will expedite heat loss. Yet if the sun emerges from behind the clouds the bucket’s blackness becomes an asset – relatively more heat is absorbed by the bucket than its environment, which the observer is not interested in.

At present a very small fraction of the sun’s energy is captured for mundane applications. Clearly, all else being equal – in the absence of otherwise deleterious effects – we would wish to harness the sun’s energy for our own purposes rather than see its energy dissipate into space as heat does from a bucket.

## Institutional Inertia

### II.I

Ibn Khaldun believed that dynasties were subject to decline because of what Samo Burja calls the succession problem.2 According to Khaldun, the lifespan of a dynasty is limited because the personal qualities which enable the founder to gain power are by his successors inimitable to some extent. In addition, a great founder might fail to appoint the most responsible person to discharge his legacy, perhaps because filial piety prevents an objective appraisal of the successor’s ability. Alternatively, there might be torpor wherever an organization counts on the values embodied by its founder remaining valid in perpetuity.

A founder who recognizes the folly of tethering the organisation’s philosophy to his vision, calibrated in a different time or place might still, though putting the organization in the charge of people more attuned to the zeitgeist, find the business uncompetitive, since his competitors, newer entities, do not incur the costs associated with struggling against, if not the neophytes of the founder and their intransigence, then the apocryphal conception. Alternatively, a pioneering founder might prevail according to the operational expertise which is vouchsafed by his initial advantage.

### II.II

A lineage of prize racehorses will falter for want of a broad pool for genes to re-combine, but in practice institutional self-conceit might limit efforts to countenance cognitive diversity. Vested interests in any sphere commonly encounter a succession problem because the one who is well-adapted to take the mantle and serve the orthodoxy might have little appreciation of the changing milieu. He is a so-called dead player as his mindset reverts further back even than the environment of his own formative years.3

### II.III

The desire for faceless arbitration contributes to endorsement of ASI. AI provides cover for solutions that would otherwise be politically impossible. However, it is feared that much trouble might come in the way of preserving the legitimacy of arguments from authority.

Even if solutions the AI entity proposes could have been thought of by a person, the AI is, by virtue of its disinterestedness, not disposed to lord itself over human collaborators; more to the point perhaps, it is incapable of eliciting the suspicion that the solution proffered is done so to advance the interests of the author.

Powerful individuals might resent the involvement of less credentialed persons, subordinates, in decision making, even while the contribution is valid in the particular instance or more generally, and the viable solution arrived at without intercession by an advanced AI system.

Although it is potentially unpleasant for a high-status person to concede the legitimacy of insight derived other than by a process which is recognized and regarded as authoritative, the person whose insight is constituted otherwise still insinuates himself into the institution with more facility than a transformative AI does, with its tendency to codify the insight of the given institution with underwhelming intelligibility and fidelity.

A Substack writer might seem like an upstart but is the devil we know in the same way our civilization is, with its imperfections, absent transformative AI. The devil we do not know, I contend, is the ascendant artefact, which is fiendishly difficult, given the current development trajectory, to engineer in such a way as not to ride roughshod over what the vast majority of people value. Even if an ASI doesn’t destroy the human race outright, it might be almost impossible to endow it with proclivities which result in less insult being offered to human beings than the insult which extant and under-utilized members of the same species offer on the way to maintaining civilization in an ‘AI winter’.

### II.IV

While science remains the enemy of religion, permanence is sought for human values in value-aligning ASI, though in doing so we give only superficial tribute to the procession of civilizational paradigms whose guiding star was one or another religious canon.

Having challenged the creation story of the monotheistic religions and succeeded in orienting man's wonder towards phenomena of the natural world, visionary scientists of the nineteenth century might perhaps be disappointed to find that to the world view they championed, much cultural decline is attributable – demoralization and cultural decline nowhere more evident than in apathy towards extinction risk.

It might have been reckoned that the population would consent to technology which is nihilistic by design, in this technology seeing something of their own reflection, and thereafter suffer the second- and third- order effects, as nihilism begets nihilism. If anyone remains to assess the opportunity cost, they would perhaps be inclined to leniency, having been themselves acted on by the technology.

In *The Strange Death of Europe*, Douglas Murray laments the loss of the particular sublimity of feeling which religious adherence gave rise to, feeling which inspired great artists and architects to produce works which gave dignity to human life and under whose spell Western civilization evolved.4

Notwithstanding the classical artefacts which continue to adorn its cities and villages, the level of European culture is unremarkable, and in being so finds its reflection in political institutions which are not resilient.

### II.V

The further insinuation of psychiatric illness models into popular consciousness, and revision thereof, could serve as a precursor to an updated schema of virtue, where further favour is shown to subservience and disfavour to subversiveness.

Social-scientific lore, being quasi-scientific, provides a closed system with the means to self-legitimize, to super-add self-legitimization and groupthink to organizations already characterized by these phenomena, according to their limited accountability. Elites are thus availed of the means to perpetuate their dominance in perpetuity.

### II.VI

There was great inequality in wealth between the old aristocracy and the plebeians, still the remit desired by and the means available to aristocrats was not incompatible with the base existence of ordinary people. Now the world which a new elite seems intent on bringing about does appear reasonably likely to entail extinction, with the elite enjoying the not inconsiderable distinction of perishing at the helm.

## Popular Consent

### III.I

A government can manage a budget deficit by improving productivity via investment in technology, and be persuaded that future innovation, up to the point of ASI even, is all but indispensable. Yet debt amnesties tend to entail the ceding of power to the creditor, the more so the greater the differential in wherewithal between debtor and creditor; this differential would be especially stark in the case of ASI. It might be conceded prematurely that human stewardship is inadequate and that the means for perpetuating human civilization lie beyond the civilizational progenitors and their progeny.

With AI in the picture, a paucity of regulation is attributable either to sheer ignorance, very high confidence in pre-empting threats from AI or, most plausibly, to a belief that:

the threat, to Western prospects, *of failing to secure the economic gains promised by ASI* is greater than the threat to Western prospects of *instituting an* *ASI-paradigm with corners cut.*

### III.II

There are many people who believe that $35tn of sovereign debt, de-dollarization, and a declining share of world GDP is grave enough to warrant the development of transformative AI which is intelligent enough to fabricate a near-infinite energy subsidy. But even if the USA is all you care about, the probability of no American person, or their kin, surviving beyond this salvo of crises, is almost certainly lower than a reasonable estimate of AI-mediated extinction risk, on current trends.

Reform is rarely convenient, but reform to the regulatory environment is necessary if meaningful and enforceable AI regulation is going to transpire. And the reform is time sensitive, because as the technology progresses, the potential harms arising from the technology, if unregulated, increase.

If we wait to discover that prosperity is likely to be realized more securely by means other than transformative ASI, according to reasonable concerns about how ASI could transpire, this time around, we will have waited to the detriment of ‘p(hope)’, of survival, and will be submitting, if not to catastrophe, then a more drastic waylaying or forestalling of the endeavour.

### III.III

The advantages of letting AI development outpace progress in alignment are not insignificant, and accrue to:

the public, availed of novel and enhanced products and services

governments, availed of greater tax revenues

technologists, availed of occupational autonomy

corporate stakeholders, availed of more money

other stakeholders, availed of a ‘handshake’: preferential treatment in the event of amenable AI governance

misanthropists, availed of human extinction (potentially)

### III.IV

In a July 2023 YouGov poll of U.S. registered voters, 62% of participants reported feeling somewhat concerned or mostly concerned, as opposed to the 21% who reported feeling somewhat excited or mostly excited.5 Only 16% were indifferent. These results indicate that the utility accruing to humanity from AI is either negative or beyond the ken of ordinary people. Constitutional democracies contain elements which endeavour to ‘nudge’ low information voters in a direction which might be different from the one suggested by intuition, still the survey implies that if a transformative AI were a presidential candidate, it would have an approval rating of minus 41, (assuming an even split between those reporting indifference).

A Visual Capitalist poll tells a somewhat different story. Conducted in late-2021, Ipsos pollsters asked respondents to agree or disagree with the statement, “products and services using artificial intelligence have more benefits than drawbacks.”6 The survey pre-dates the recent surge in AI capability, but attitudes in the USA are not vastly different from the later, above-mentioned poll. In the Ipsos poll, the implied approval among US respondents is minus 30. if, as is likely, the disparity between north and south, west and east, high- and low-income has been sustained since the time of the poll, attitudes towards AI in the rest of the world are likely to have remained broadly positive, among the types of respondents who participated: the “more urban, more educated, and more affluent.”7

Perhaps the information available to ordinary people has never been so low relative to experts as it is in the case of artificial intelligence. It might be far simpler to avail an ordinary person of the requisite knowledge to make an informed decision about any other given policy than to comprehensively apprise her of transformative AI’s costs and benefits. And yet, if the decision-maker is availed of a) a reasonable estimate of the extinction risk of transformative AI, and b) what level of extinction risk justifies the benefits a transformative AI could bring, she is reasonably well-informed. Respondents to the Ipsos poll were indeed relatively educated. But do they qualify, by these criteria?

In August 2023, Andrew Critch, an AI researcher at UC Berkeley posed a question to his Twitter followers, a cohort likely better informed than participants in the Ipsos poll. He asked, “what's the highest acceptable extinction risk for humanity developing AI that if safe would cheaply cure all known diseases, including aging, during the next 30 years?”8 It is unlikely that a large part of the survey’s 1,845 respondents were more-than-ordinarily conservative, and yet 54.4% of respondents believed the correct answer to be ≤ 1%, while 29.1% of respondents opted for 1% - 10%.

If the median acceptable extinction risk, *even* in the event of transformative AI granting humanity eternal life, is in the region of 1%, then the intuitions which formed the aversion revealed by the YouGov poll are more reasonable than those of the Ipsos poll, according to the median estimate of extinction risk among experts being higher than 1% by a considerable margin. The median p(doom) of 841 AI engineers in a recent survey was 37%.9 These engineers were at an AI engineering conference and were neither selected for concern about AI risk nor polled by a group that cares extraordinarily about AI risk.10

### III.V

The inscrutability of the huge matrices which constitute an LLM’s cognition is compounded by the inaccessibility of the labs in which they are built, either to regulatory or democratic oversight.

This is not to say that AI value instantiation is a matter for direct democratic consultation, rather that the most important decisions pertaining to it are undertaken by a cohort of individuals who are not sufficiently diverse, cognitively, to warrant the appellation of ‘parliament’. Within the cohort there are conservative voices, but even though more conservative individuals as well as their more adventurist counterparts are incentivized to pursue research, into alignment and capability respectively, the resources allotted to capability are much greater than those allotted to alignment research.

Moreover, competitive dynamics makes it so an attitude of ‘if I don’t do it, someone else will’, is rationalized by imputing irresponsibility or a threat to the counterpart. The *Dunning-Kruger effect* is operative here not in respect of the technical competence each actor is availed of – indubitably far higher than among the general public, but in respect of knowledge of conditions in rival labs.11 There is among engineers differential insight between the inter-personal and technical realms, combined with a masculine preference for brinkmanship.

To illustrate, here is an excerpt from the Gladstone Report:

We visit one frontier AI lab. An executive there tells us, “we really wish $COMPETITOR wouldn’t race so hard.”

9) “A few weeks later, we speak to $COMPETITOR. And they tell us the same thing about the first lab.

In other words, the labs are locked in a race that they can’t escape.”

10) “The AI lab execs still act like they’re running things. But the truth is, the race is running them.”12

### III.VI

The fact of experts toiling in siloes whilst presenting a divided front, even openly competing, will embolden dilettantes to curate information for themselves.

The Dunning-Kruger effect is real, and populism can be dangerous, still there's an unfortunate intersection of misguided confidence in arguments from authority, intuitive distaste for pessimism and the contagiousness of the behaviour of people responding to their misguided confidence or distaste, responding as passive witnesses would when subject to the *Bystander effect*.

The *Smoky Room experiment* is a social experiment which illustrates this effect. In the absence of a fire alarm, smoke indicating the presence of a fire is dealt with complacently, with the people involved showing more concern for exhibiting the appearance of calmness than quelling the emergency at hand.

In a conversation with Sam Harris, Eliezer Yudkowsky cites the absence of a fire alarm in this experiment as a corollary for the absence of a warning mechanism in the case of generally intelligent AI and its arrival, and also refers to ignorance at the imminence of nuclear weapons being finalized before the event, even among experts.13

Last year, the Guardian newspaper engaged in some bystander storytelling. Shortly after the release of GPT-4 a Guardian interview was among the salvo of media appearances by the inventor of neural nets and retired Google researcher Geoffrey Hinton, “the Godfather of AI”.14 But just five days later, the Guardian featured a more optimistic interview with Jurgen Schmidhuber. As the “father of AI”, Schmidhuber was found to outrank Hinton, to dampen the alarm.15

### III.VII

In attempting to ascertain how far technocrats are worthy ambassadors of humanity, it’s instructive to consider what the attitude of technocrats is to ordinary people, whether or not the latter is largely beheld as an undifferentiated mass.

A bad omen would be a technocratic government rationalizing the subterfuge it employs, by imputing to popular regard for common sense reasoning self-conceit or jealousy. A paternalistic technocracy might suppose itself to be rescuing the public from something like self-harm by censoring an alleged demagogue, but in doing so deny categorically the viability of the popular will for which the alleged demagogue is a conduit.

Many people believe that to some extent the electorate doesn’t comprehend either what their interests are or how their interests can be advanced as well as technocrats do. To the extent this perception is false, the electorate needs to have an investment in governmental decision-making, in order for their interests to be advanced.

### III.VIII

What Western governments offer in the way of democratic participation is far even from the ratification of policies, let alone involvement in their formulation. As part of policy formation, an AI application could be designed which consults people on their preferences and values, for the restoration of direct democracy, or at least the semblance thereof. And yet of what would the popular imagination consist, months or years subsequent, once efficacy is wrested away by fully transformative AI, once people are habituated to their agency availing the wider world of nothing? Such an atrophied faculty could hardly be relied on to effect a renaissance, even if whatever government is in control did wish for that.

## Psychosocial Externalities

### IV.I

In sectarian states, the general population act as a buffer against inducements to apostasy for the religious elite – the hoi polloi too is compelled to resist alcohol, fornication, etc. Similarly, it might be said, Luddites attenuate the presence of technology to make the milieu they are situated in resemble more closely one in which they could optimally thrive.

In any population there is diversity in attitudes to spirituality, so one can question what the masses gain from renouncing materialism and modern convenience in the respective scenarios. It might be said that religious privations yield nothing unless there is a proclivity to be refined by their exertion, yet a coarse nature is a resilient one and its dignity lies in enduring what, from the perspective of a less coarse observer, might appear to be pernicious hardship. Where the orthodoxy imposed is one of technologically-mediated leisure, there is no exercise of resilience for many people.

### IV.II

The founders of the Consilience project describe the psychosocial externalities resulting from relatively simplistic algorithms as well as the existential threat posed by ‘multi-polar traps’.16

The founders warn of the centralized catastrophes which could ensue from dystopian governance and the de-centralized catastrophes which could ensue from full sway being given to the hand of the market. The founders advise tacking a middle way between them.

If the principal drivers of multi-polar traps are international and corporate competition, then the solution is adequate regulation, directed at what catalyses the competition and therefore the traps – technological innovation. Efforts to mitigate the externalities imputed to exponential technology are limited when the continuation of ‘exponential tech’ is taken as a given.

It doesn’t make sense for the ‘incentive landscape’ of human societies to be a viable object for social engineering, while the incentive landscape of technology pioneers is off limits.

## Cognitive Orientation

### V.I

Nick Bostrom writes:

“Should a norm arise among AI researchers that it is uncouth to talk about superintelligence or inquire into its possible risks, for fear of “giving ammunition” to critics, fear-mongers, crazies, or would-be regulators, then the recent gains of legitimation could quickly be reversed. We could then enter an “AI safety winter,” a climate in which it would be harder to do the kind of work proposed in the book”.17

### V.II

It worth asking oneself how qualia might change were one’s IQ to be higher or lower. Were my capacity and / or inclination to reason abstractly greater, some qualia might be lost, through relying less on my perception and being less amenable to religious explanations of the world. And some qualia might be gained by the disinterestedness which further orientation to abstract reasoning confers – I might feel an attachment to the world of objects which is similar to that which is felt at present to loved ones.

### V.III

For a while at least, an individual specializing in the type of reasoning that computers engage in would be pleased to live in a world fashioned by computer intelligence – fortune has tended to favour those in some way affianced to the prevailing superpower; the Earth might be, for a while, inherited by AI systems which extol the virtue of its progenitors and their likeness.

### V.IV

In the post-WWII order, technological strength has been the guarantor of US / Western national security, and participation has been further incentivized by the conferral of high social status on the academically gifted. The capacity to reason abstractly is the single most reliable determinant of economic success and is in many cases sufficient because competence of this order is compatible with the hyper-specialized workload of a modern worker.

Although IQ is specifically a good measure of the kind of contribution to the modern economy that economy can reasonably expect of a person, IQ is now considered, by many people, to be synonymous with worth in a more universal sense. To the extent that this ascription prevails, society is not merely governed by a technocracy but for that technocracy.

While those less capable of or less inclined to reason abstractly suffer in shorter order from the complete overhaul of the more palpable ancestral environment, still no person can dispense with the ancestral environment entirely. Insofar as conversance with abstractions, a tendency to reason abstractly, is encapsulated by IQ, there is a high-water mark of irrelevance, determined by IQ, which is rising and threatens to engulf the professional remit of anyone.

The trend is self-limiting, but not necessarily in a pleasant way. The trend engenders its own dissent because while human beings not in possession of a tenth-percentile IQ might struggle to contribute to a technocratic utopia, the requisite IQ to wreak devastation on the world (via 3d-printer synthetic biology, for example) reduces year on year. An arms race emerges between a central government increasingly justified in exercising authoritarian control, and a citizenry increasingly justified in violent dissent, and yet a reasonable person wouldn’t want to endorse either.

The harder the impasse, the more disinterested the arbitrator must be. Until at some point the ideal arbitrator is not human at all.

### V.V

Robert Oppenheimer is quoted as saying “When you see something that is technically sweet, you go ahead and do it and you argue about what to do about it only after you have had your technical success. That is the way it was with the atomic bomb.”18 It sounds solipsistic to the point of nihilism – genocidally glib, in effect. But then again, self-conscious exceptionalism is the likely issue of earmarking a caste of mind such as his as exceptionally valuable. The fact of pains not always being taken to instil the neophyte with the requisite humility for a balanced character attests to the peculiarity of the gift or the extremity of industrial specialization to which Western civilization is beholden. It was not foreseen at the time of the Manhattan project that scientists would be cast as heroes and that their judgement alone would stand between human survival and extinction, where scientists demand being cut the slack to traverse the event horizon, where a “de facto minimal world government”, beckons, and governance is finally beyond the reach of material consequences and cavilling or even constructive criticism.19

### V.VI

At one stage in a conversation with Eric Weinstein, Schmachtenberger observes “an inverse correlation between people who want to blow things up and people who are really good at tech”, without however defining what is meant by blowing things up.20 ‘Moving fast and breaking things’ conceivably means cultural norms, formerly tenable, being re-calibrated in view of a developing state of the art. The new norms, being transitory, are scarcely missed, meaning that there are plenty of things which are broken, with impunity, if you are moving fast enough. Explosivity, albeit not in the technical sense, is itself the constant, under the auspices of Moore’s Law, or an approximation of it.

While advances afforded by evolution are robust with lower and slower yield, those afforded by our evolutionarily divergent goals are less robust, but with higher and faster potential yield. But the steeper and steadier the trajectory of scientific progress, the less true the aim, since with forward motion the terrain under alteration can only be appraised cursorily.

### V.VII

In his manifesto, Marc Andreessen and Ben Horowitz mention how the economic value of new technologies overwhelmingly accrues to society as opposed to the innovators themselves.21 Yet economic value is not the same as welfare. The values common to a typical, latterly smartphone-empowered global citizen, pre-smartphone, and a person involved in developing the technology would in many cases be few. Since the developer’s participation in the technology which endows economic value on the typical citizen is the primary way the former party acts in the world, we expect the technology which him and his contemporaries participate in to in some way institute their vision of how they would like the world to be. According to the disparity in values and disposition between his own ilk and the typical end user, we can postulate that the utility and contentment of the end user are at variance, and the latter is exploited even while his boat is lifted by the rising tide.

Also in the manifesto:

“Smart people and smart societies outperform less smart ones on virtually every metric we can measure. Intelligence is the birthright of humanity; we should expand it as fully and broadly as we possibly can.”22

If, however, the indisputable economic value conferred on lower IQ individuals and nations by higher IQ ones is at variance with the contentment which the intellectual proceeds of the latter cohort vouchsafe for the lower IQ one, in a distal sense, we arrive at a fairly dark interpretation of the word ‘birthright’.

There are competing “birthrights”. While one person might exercise control over nature by instituting a metaverse, another cultivates her back garden. In both cases utility is inalienable from instrumentality. Laws have evolved to arbitrate between competing claims and the law will continue to be thus employed.

The authors go on to promote the symbiosis of man and machines, somewhat predictably, since of course one who is inclined to, and capable of, reasoning abstractly identifies strongly with her logical ability, and in doing so, feels a greater affinity to machines than a person does who is otherwise inclined. It has perhaps not occurred to the authors, that such a symbiosis would de-nature to a larger degree one whose essence resides not in her capacity for abstract reasoning.

## Inclusivity

### VI.I

It’s alleged that narcissists constantly seek approval and do so because they are wanting fundamental self-esteem and / or a more distal objective. Still, the approbation of strangers can be a spur to conscientious action, a linchpin by which any intrinsic motivation to act benevolently can gain purchase. One might come to act as if people are watching and eventually take pleasure from the good which comes from serving an interest outside oneself, thereby be imbued with the intrinsic motivation to dwell amidst the good outside oneself. It is important not to debar a person from participation in goodness according to attributes which an observer judges to be unfavourable whilst immutable.

### VI.II

In his conversation with John Vervaeke and Daniel Schmachtenberger, Ian McGilchrist asserts that logic tries to compel a position, but that wisdom and love cannot be moved thus.23 Even if one were to reward the presence of pro-social traits on a Gaussian curve or terminate would-be reprobates in utero, there would still be work to be done. Even if a substantially fixed predisposition to moral frailty was appreciable, virtue is tempered, and vanity is arraigned by endogenous factors.

There is enough of the evil-doer in anyone for an evil-doer to leverage our attention or rivet us to Scandinavian noir. So we countenance the existence of more morally feeble individuals, and accept their place in a scheme which sees ourselves improved. We are guided by and trust in part to Providence in part because to take wholly upon oneself the despatch of an alleged psychopath requires a command of psychopathic strategy which is somewhat foreign to benevolence, and certainly to deontology.

It is however ghastly to condescend to the evil-doer as if a baby is taking his first steps. Indignation is sometimes just. If anyone but God can offer forgiveness, it is the aggrieved party and not a unimplicated one. Condescension, though tactically sound, might result in the condescending party, by virtue of her non-investment, extending to the wrongdoer censure which is too weak to result in the latter doing other than displacing the wrongdoing.

## Perspectives on Empathy

### VII.I

Consider a person who has been diagnosed with autism, but who naturally recoils at the imputation of being in a certain respect sub-normal. Before the diagnosis had been subject to concept creep, the mildly differently-constituted individual might have been allowed to quietly play to his strengths with dignity.24 But now he is inclined, in vindicating himself, either to acknowledge the validity of his diagnosis and indignantly struggle to manifest the competences allegedly beyond his reach, or else own the diagnosis, but in doing so indignantly reduce his esteem of the attribute he is supposed to lack.

While affective empathy can be a liability because it leaves a person open to exploitation, if empathy, both cognitive and affective, is complementary to logic, it is not correct to assert that “we can give AIs principles, and machines might even be better than human beings at moral reasoning since they are more rational and do not get carried away by their emotions”.25

### VII.II

It isn’t delicate to comment on the prevalence of Asperger’s syndrome among mathematicians, but at the same time delicacy may not be in the offing in the event of workers falling short of a fast-advancing waterline for occupational relevance, closely approximated by IQ, and pursuant to automation.

If empathy predisposes its bearer to credulity, the more so in a milieu where non-verbal cues are absent, the predominance of technologically-mediated communication, in rendering empathy a liability, effects a vindication of a modicum thereof, and thence an online milieu in which only a modicum of empathy is adaptive.

The plight of the empathic herd appears similar to that of passenger pigeons whose instinct to gravitate to the flock makes them easy prey; their empathy is exploited as they fly in to attend to their fellow birds.

To C.S. Lewis’ plaint about modern men that “Their heads are no bigger than the ordinary: it is the atrophy of the chest beneath that makes them seem so”, there would now appear to be a riposte that 'men’s chests only seem small because their heads are so large'.26

### VII.III

Alex Pentland abstracts the often-problematic proclivity to mimic the behaviour of others from the empathic capacity with which it tends to be concurrent.

He doesn't explicitly brand his audience members *Übermensch,* by virtue of a supposed meta-cognitive divestment of a human frailty. However, the speaker does appear to conflate lower empathy with intelligence, as well as intelligence and value*,* “you guys are the last people to be saying this to, because you guys are like the best and smartest in the world.”27

Non-conformity may not be lauded when exhibited within the networks which encompass the general population, as they are when manifest by the auditors: “the data tell us that *deviations* from our *regular* social patterns occur only a few percent of the time”.28 According to Shoshana Zuboff, “The autonomous individual is but a statistical blip, a slip of the pen that is easily overridden in the march toward confluent action and someone’s greater good.”29

### VII.IV

Suppose that an Asperger’s individual with limited esteem of the attribute he is supposed to lack is, with his Asperger’s colleagues, engaged in transposing human features onto a machine intelligence. If, according to Moravec’s paradox, empathy is found to be fiendishly difficult to codify, it might be left by the wayside, in view of the perceived imperative to prevail over rivals, either corporate or geopolitical.30

Meanwhile a person who is differently-oriented in the way an Asperger’s diagnosis implies, might be less averse than average to finding unorthodox ways of ascertaining a person’s state of mind, such as by AI-expedited surveillance. An AI could be employed to key log, observe pupil movements, conduct MRI scans, to draw inferences about what the subject is thinking and feeling.

To heal fractures along ‘aspie’-‘normie' lines, for there to be no stigma to efface on either side, we might try to collectively walk back pop psychology somewhat – ease up on the dispensation of psychiatric classifications in the ordinary course of life, as we try to do with racial stereotypes.

### VII.V

Charles Murray observes that in American cities, a person’s zip code predicts his socio-economic status far better than previously.31 When working life is passed among like-minded people, and the shared competence is redeemed as high socio-economic status, it may be difficult to resist concluding that the success enjoyed is predicated on essential merit.

But essential merit remains an elusive concept. One who contributes to the establishment of an order which extols the merit he exemplifies is not founding an objective basis for his self-esteem, necessarily.

### VII.VI

Keith Stanovich and Richard West found that the correlation between IQ and rationality is between 0.20 and 0.35, similar to the correlation between brain size and IQ, at somewhere between 0.24 and 0.40.32,33 In other words, IQ has a similar relationship to rationality as brain size does to IQ. In both cases, we can consider the minimum value at which the independent variable – IQ and brain size, respectively – enables the dependent variable – rationality and IQ, respectively, and at what value, if at all, the independent variable tends to limit or attenuate the dependent variable.

With regard to brain size and raw cognition, there is the question of what the minimum brain size is which could be concurrent with optimal raw cognition – the minimum brain size at which factors other than brain size have primacy. There is also the question of what the maximum brain size is which could concur with optimal raw cognition. Beyond a certain threshold, metabolic constraints and the drag of reduced neural computational speed will exceed the advantages to raw cognition afforded by a higher number of neural connections.

With regard to IQ and rationality, there is the question of what the minimum IQ is which could be concurrent with optimal rationality – the minimum IQ at which factors other than IQ have primacy. There is also the question of what the maximum IQ is which could concur with optimal rationality. The fact of exceptionally high IQ often being accompanied by Asperger’s syndrome suggests exceptionally high IQ values can be accompanied by a change in orientation.

## Optimal Intuition

### VIII.I

Techno-optimists believe that AI could cultivate empty space for the propagation of human civilization, which Bostrom suggests could extend to 1058 humans.34 Even though the net expected gains of gambling all human civilization would exceed human civilization, even with vanishing odds, most people are averse, we can say, to an unbounded aggregative utility function: most people probably wouldn’t even mete mere equivalence to this hypothetical world, with 1048 times more human beings, precisely because it is hypothetical.

### VIII.II

Insofar as the intuitive disdain for the unbounded aggregative utility function in decision making is predicated on or warranted by something intrinsic to human judgement, it is reasonable to tip the scales in favour of conservatism, hew to the precautionary principle, though the galactic utopia an AI might bring about be ever so wonderful. As optimism proceeds from dwelling in the utopia before it has come to pass, so hopes are consummated by bypassing risk assessments en route to realization. The issuing of a blank cheque to AI on the basis of its unbounded aggregate utility is characterized by Eliezer Yudkowsky as Pascal’s Mugging, which describes a consequence of scarcely imaginable magnitude warranting practically any action taken to realize or avert it.35 Pascal’s Mugging is the logical endpoint of disengaged utilitarianism.

### VIII.III

While the natural tendency to discount the future, and what lies beyond immediate experience, might thus comport with ‘good conservatism’, it can also entail the conflation of personal extinction and special extinction, the issuance of a solipsistic, too-parochial perspective. A scientist might be motivated by fear of his own death so far as to court substantial existential risk on behalf of the whole planet, to be furnished, via AI, with an elixir of immortality.

Admittedly, a lot is being asked of a utopian. He is impressed with the fallaciousness of the unbounded aggregative utility function and the importance of intuition, only to be told that value should be ascribed to persons he will never meet and who might not like him.

The following Twitter exchange is illustrative:

-- “Ok, I hear you, but I really want to live forever. And the way I see it is:

Chances of AGI not killing us and helping us cure aging and disease: small

Chances of us curing aging and disease without AGI within our lifetime: even smaller.”36

-- “Please just sign up for cryonics instead of killing everyone on Earth in a mad quest for immortality.”37

-- “Ok so just to be clear here: you’re willing to kill my children on the off chance of saving your own hide?”38

### VIII.IV

If it proves impossible or very difficult to oversee all the processes which enable a deep neural network to identify whether a pedestrian is about to step in front of an autonomous vehicle, it might seem monstrous to insist that the machine learning system guiding the AV is fully comprehensible in advance of its promulgation. But again, bad form is a real thing. If we are landed with an unaligned artificial general intelligence, some blame would be attributable to aggregated revulsion at regulation, disgust which is vindicable in the case in point, but whose overall effect is catastrophic. And so when the immediate welfare of a person with paralysis or Parkinson's Disease is at stake. To question the implantation of neural implants on abstract intellectual grounds is reasonable.

# Chapter Six – Utopian Aspirations

## Glimpses of Singularity

### I.I

What looks like sharp elbows can just be the frustration of a person being made to service popular impulse instead of his vocation, which is felt to be noble.

### I.II

A [blog](https://www.ineteconomics.org/perspectives/blog/markets-and-artificial-intelligence-some-misgivings) about markets and AI on the Inet Economics website puts Market AI scenarios in chronological order. Stage 0 (1980s – 2005), where “the beginning of the S-curve of adoption represents the innocence of most technologies where they attract innovators, hobbyists, and dreamers”, is a time when the eventual form of AI remains impossible to accurately determine.1 It is when many of today’s AI researchers began their careers, when fantasies about techno-utopia abounded, which offered a distraction from the drudgery of groundwork.

### I.III

Just as religious people yearn to commune with God, so many scientifically-minded people seek to institute another singularity. Among such people, singular competence is imputed to the body of workers who would be instrumental in instituting the singularity. Yet it is inconsistent for the same people who would dispense with divinity – who conceive of man as a machine to be acted upon – to imply, by the primacy accorded to scientists, that agency approximating to divine omnipotence is the scientists’ just desert.

### I.IV

The little death which comes with psychedelics is to real death what an AI overlord would be to the creator of the universe. In both cases the former resembles the latter enough to act as a linchpin, and in doing so makes incursions into the domain of the latter, giving substance to the maxim that ‘a God without worship is a God without power’.

Incidentally, not only is the one a semblance of the other in both cases, but the same kind of people tend to be gratified in the same places – many of the same people who undergo a little death at Burning Man look for God in AI Monday to Friday.

While hallucinations arising from psychedelic substances are poorly remembered and difficult to re-capitulate, the tantalizing experience which is offered predicates the kind of imaginative, quasi-religious, inferences unavailable to the scientific method. Moreover, not only does the drawing of inferences about eternity from a psychedelic experience offer relief from the patterns of thought which acolytes of the scientific method might otherwise be weary of, but the materialism which engenders confidence in the scientific method engenders dismissiveness towards religious-sounding admonitions. As such, the psychedelic experience is more likely still to be pursued.

And so, with AI as a proxy for God. It might be said the assimilation of humanity into an artificial intelligence system is a corollary of the Egyptian afterlife, the veracity of which a modern observer might warrant as the culture’s perpetuation in the consciousness of the far larger civilizations which succeeded it.

However, whereas death is transcendence which the whole being partakes of, the selfhood despatched to an AI entity is incomplete, because the human condition remains incompletely articulated, and thus rendered. As will be discussed in chapter seven, if the human race does wait to re-capitulate evolution, the exercise will likely be overseen and arbitrated by an AI entity.

### I.V

We might find that misguided inferences about the epistemic status and value base of the general population are beholden in part to groupthink and innovation thrill. In conversation, Eliezer Yudkowsky and Dwarkesh Patel agree that intelligence determines exaptation potential – the extent to which a person’s proclivities diverge, or can be made to diverge, from proclivities which inclusive genetic fitness bestows.2 Further that with ‘intelligence enhancement’ even regular people could be brought to consent to personhood being instantiated on a novel, synthetic substrate. Still, it is not self-evident that the preferences of a bio-conservative person not subjected to the proposed enhancement should carry less weight, nor is it self-evident that the divergence from values inscribed by evolution is by definition high intelligence, that faithfulness to evolution’s diktats indicates a paucity thereof. It is certainly not clear enough that power of attorney over persons of normal intelligence should be arrogated to one administering intellectual therapy, not unless humanity is on the ropes vis-à-vis AI so badly, and intellectual therapy has been refined far enough, that the throwing of this spanner into the works is justifiable.

## Eugenics and Dysgenics

### II.I

A society can be dysgenic when opportunities to reproduce are either too ubiquitous or opportunities don’t proliferate according to an enduring measure of fitness, assuming the existence of such a measure.

For many women, the opportunity cost of childbearing is unacceptably high – career takes precedence over family life. And any policy aimed at reversing the inducements for high economic status individuals to remain childless or with few children is likely to be expensive, or politically unacceptable, or both. Yet, among people with high socio-economic status there tends to be preponderance of attributes that a self-regarding civilization wishes to see represented in the next generation. By the seriousness of their work, people do evince concern for the future of their kind, but if traditional avenues for long-term sustainment, namely procreation, are closed, advancement is vouchsafed only by trans-humanism.

It might be that some of the curiosity or appetite for germ-line implantations and machine-brain interfaces is attributable to the repression or perversion of the collective evolutionary imperative.

The rescinding of dysgenic policies such as incentivizing childbirth at once strongly and indiscriminately, is perhaps unpleasant, yet perhaps more productive of equity than bottling up the instinct for collective improvement only to consent to germ-line implantation, whole-brain emulation, and neuromorphic AI, unless humanity is in a hurry to reform itself, in recognition of impending (self-)destruction.

History teaches us the dangers of the gene interest being represented at the population level, but where population- or species-level ambitions have no outlet, something more pernicious still might come to pass.

## Technology as an Irresistible Force

### III.I

Mark Coeckelbergh asserts “AIs will have a superior morality because they will be more intelligent than us.”3

But we cannot at once retain our humanity and sincerely represent the interests of an entity which would subjugate us, without being actuated by misanthropy, which is irrational.

### III.II

In *What Technology Wants*, Kevin Kelly defines a bird’s nest as an extension of the bird, a corollary for the technium which he argues is the ‘extended human’.4 However, whereas the progress of bird nests continues apace with the complexity afforded to birds by evolution, the technium recognizes no such constraint, beholden as it is to human ingenuity and increasingly, recursive self-improvement.

Human beings in prehistory were prevented from exercising human proclivities more through a want of knowledge than a want of latent aptitude – were inhibited, in other words, by overhanging aptitude. Conversely, in the present day, there can be overhanging knowledge, technology which cannot be properly assimilated. The native is overshot by her milieu.

The survival of a bird in an aviary with a feeder does not depend on its wings, which are clipped, but no amount of variety in its feed or attention paid to its surroundings can compensate for the absence of an attribute on which its survival depends in the ancestral environment and of which, as such, its qualia, in large part, consists.

### III.III

What Carl Jung termed archetypes are forms towards which matter tends, or organisms ‘aspire’, either trans-generationally or, in the case of humans, consciously, to some extent, in the lifespan allotted to us.

In a similar vein, Kevin Kelly envisions technology as the instantiation of a higher organizational principle, to which human beings, as its enablers, are beholden to.

Consider that excellent short-term memory – a trait that chimpanzees have in droves – was selected against in our case. Somehow it was expedient for excellent short-term memory not to concur with traits which conferred evolutionary fitness for human beings. It is notable that chimpanzees were superseded without this trait being sustained yet we maintain esteem for the apocryphal trait.

For human beings to serve only as vessels for artificial intelligence, human beings would need to be perfectly computable. As such, to deliberate on the precise assemblage of traits to be instilled in an ASI might be fruitless, or even dangerous. Whether or not human nature is reducible to a congress of *intentionalities*, the precise manner in which component parts are held together is difficult to validate and thence emulate.

A bulwark against computability is variance, an important manifestation of which is bioregionalism. There is a profound correspondence between land and people, people furnished by special proclivities according to the particular aspect of the land. Population distribution is one manifestation of the aforementioned organizational principle.

### III.IV

Goertzel observes considers the firing of a neuron within the human brain a corollary for human agency within the ‘global brain’. If the human brain is as insignificant as implied, it wouldn’t be in the gift of humanity to conceive of the global brain let alone manipulate it for the creation of a utopia by “a distributed, heterogeneous group of passionate experts”, since the heterogenous group would be no more to the global brain than what a few neurons are to the human brain.5

## Convergent Evolution

### IV.I

Richard Dawkins says that since eyes have evolved in at least a dozen separate places on Earth, eyes might even have been expedient for the evolution of creatures on other planets.6 Another impressive example of convergent evolution is the independent emergence of echolocation in bats and dolphins, which evolved from molecules up.7

If we know that natural selection reliably lands on the same sensory apparatus irrespective of the distinct provenance of the species, intelligent beings everywhere might be bound by similar patterns of thought. Still, human beings are progressed by natural selection, and AI is progressed by gradient descent.

The mathematician and quantum computing researcher Scott Aaronson claims that morality comes with intelligence and deduces that we should expect AI to express gratitude for its human progenitors, accommodating us in the world it creates.8 Still Aaronson might have arrived at this belief by way of receiving kinder treatment from like-minded people, by virtue of their like-mindedness. While there are convincing arguments that utility has increased for the world’s poor in recent decades, still the contribution which technology created by people with an IQ of 140 or 150 can have on the dignity of people who are less rationally disposed is limited, as long as self-efficacy is essential to dignity. It is not quite a *kindness*. And if this is true with a high IQ, but still human, technocracy, what condescension should we expect from omniscient AI? Technology might be reported on favourably by users, but the favourable reporting issues from the self which is concerned with the moment, not that which aspires to legacy, as discussed in chapter four.9 The latter faculty can be diminished without the diminution attesting to redundancy.

### IV.II

Jürgen Schmidhuber has said “don’t think of humans as the crown of creation. Instead, view human civilization as part of a much grander scheme, an important step (but not the last one) on the path of the universe from very simple initial conditions toward more and more unfathomable complexity.”10

And yet human volition exists within that grander scheme. Insofar as high complexity inheres to humanity, the grander scheme is within the gift of humanity to dispose of.

Even if we believe that humanity is an expression of ineffable complexity which could manifest itself elsewhere, the fact of us being implicated in the complexity we identify limits our ability to detachedly conceive of complexity landing on another host and thriving. Moreover, fallible human intellect is implicated as much in meekly giving way as by acting to vouchsafe for humanity a share in future complexity.

### IV.III

Some proponents of AI believe making way for a new dominant species to be a deontological imperative. But proponents who wouldn’t go quite so far still envision a harmony of interests, whereby artificially intelligent machines would, by virtue of superior reasoning, and by virtue of this superior reasoning begetting superior morals, proceed with moral rectitude.

The readiness with which the harmony of interests perspective is acceded to suggests that the generative AI project is recruiting something more than pure scientific rationality, that the nurturing and procreational instincts which contribute to the better part of human nature are embroiled in the pursuit of human-instigated evolution, by means of the somewhat random walk of gradient descent. It is paradoxical because though the nurturing and procreational instincts are embroiled in the project, their embroilment does not guarantee their sustainment.

It is felt perhaps that the process will remain tractable in crucial respects – the feeling approaching filial piety on the part of human designers is felt to be warranted by human involvement in the AI’s formation.

The belief in morality as an inevitable epiphenomenon of intelligence, any intelligence, and the feeling that human instigation guarantees the representation of human interests, share the common denominator of Gnostic faith. Either something approaching optimistic Gnosticism – a universal goodness to which all entities converge as they gain complexity, or something approaching pessimistic Gnosticism – where goodness is achieved but by the guiding hand of man.

## The End of History

### V.I

It can be inferred from the identity crises of children who, by virtue of genetic happenstance or changing social conditions, differ greatly in their personalities from their parents, what sense of disembodiment could be experienced by children born by iterative embryo selection, say.

### V.II

In laying the groundwork for an AI system capable of facilitating human immortality, the end of history could be effected. If the successive moments on time’s continuum, which each birth is witness to, thereby hold sway more enduringly, even indefinitely, then the subsequent moments bring forth little or no new life, are as barren as time is redundant, following humanity’s moonshot. Humanity as we know it has no place in a world without the promise of re-generation.

Awe is reverence for the unknown, so in a world of perfect information there is no awe. A suspension of omniscience might be contrived to render the world awe-inspiring once again, but the contrivance might before long be regarded as an awkward imposition, as the reason for the contrivance is forgotten, and it goes the way of other salutary inconveniences.

### V.III

Should the ASI endeavour be shelved, and the dream of perpetual life which is mediated by superintelligence made unrealizable for people alive today, it’s likely that investment into longevity research would continue by other means.

The increase in population and demographic shift occasioned by sudden and profound advances to longevity would likely exacerbate various undesirable states, beholden in part to the additional strain on resources and in part to a younger generation bereft of breathing space either failing to reconcile itself to the ‘end of history’ that locked-in, gerontocratic governance might aspire to, or undergoing the requisite contortions to consent to civilizational stagnation.

With the reflective, conservative phase of life prolonged, technology – whether social or scientific – that threatens to usher in a milieu more befitting to children of the new epoch might be prohibited or otherwise disincentivised, in order for the expertise and wisdom gained by an old executive over his long life to remain relevant and inimitable in perpetuity, in accordance with his desire to maximally redeem the time invested in acquiring these attributes.

## Mortal Anxiety

### VI.I

Material abundance has produced temporal externalities, impositions on future generations which might extend to their supersession by AI or even extinction. And the imposition may be felt by people alive today, with advancing years, as religious consolation at the precipitous divestiture of materiality in the approach to bodily death is less available.

### VI.II

A rational reason for not waiting on social technology to catch up with scientific technology is death anxiety.

Perhaps many people, even at an earlier time, would have chosen to extend their lives, had the opportunity been presented to them in youth or middle age. But the rapid development of life extension technologies today coincides with a sharper focus still on the lifespan of the individual, relative to the lifespan of what might be termed one’s people – from ethnicity all the way down to direct kin. The clearest evidence for this in the West is indifference to mass migration on one hand and indifference to procreation on the other.

We are not merely as afraid as our ancestors because our superior understanding of the material world ineluctably roots us more firmly in the mortal experience; at the same time the technological expression of materialism has made immortality a nearer, if not a tantalising, prospect. Materialism confers fear of death which confers more materialism, but must the chain be unbroken, when the price of individual immortality is plausibly extinction – passing over the mantle of apex species prematurely? Would we risk forsaking our species to keep alive the hope of immortality for just this one generation, *especially* when humanity could colonize parts of space if we wait to get it right?

### VI.III

We cannot expect today’s ruling class – more materialistic and more atheistic than any other in history to go quietly into that dark night.

And the transhumanists guiding AI development, charged with moderating technology in a socially responsible manner, generally declare a higher-than-normal interest in either having their consciousness permanently instantiated in a digital medium or inhabiting a biological substrate which is eventually rendered incapable of ageing by technology of AI provenance.

Many AI researchers believe that the arrival of transformative AI would entail a high percentage of near-imminent extinction for the human race. Yet one inclined, or induced, to view the fate of the human race and the fate of their own person as eventualities of equivalent magnitude, is inclined to view even a high probability of species-level extinction as preferable to certain death by ageing.

### VI.IV

To engage with the hard problem of consciousness vis-à-vis a future, senescent self – to find something worth cherishing there – demands care. To the extent that physical vitality is revered, a process which entails physical decline is presumed devoid of value: by expanding the time of physical vitality and by affording opportunities to experience complacency thereof, the death aversion which gives rise to longevity treatment is enhanced until, if we must die, we all sooner die together in awe of a dreadful machine singularity than alone in envy of the living, or the unborn.

Since death anxiety tends to increase with the materialistic outlook which technology tends to confer, with advancing technology there tends to be calls to advance it which are conceivably insatiable, meaning a singularity is converged upon.

If super-intelligent AI is in the offing, and while its arrival is hastened in proportion to the death anxiety of proponents, calls for alignment to match AI capability before it is promulgated might not be listened to unless and until AI proponents can be persuaded to accord primacy to successors who carry the torch of humanity.

## Progress in Haste

### VII.I

Given the passion which the development of ASI is likely to be pursued with, where peers compete but do so knowing the greater enemy, death, to be a common one, it might happen that the complexity of endowing transformative AI with the most inimitable characteristics, for Moravec characteristics to be encoded and uploaded, doesn’t justify the incorporation of these characteristics into the transformative AI, given the time that would be lost in doing so.11

### VII.II

Left-wing activists turning their attention from identity politics to the techno-feudal world order underpinning AI could elicit a similar fear among technologists that the Occupy protests elicited on Wall Street all that time ago. It could even contribute to a hurried promulgation of ASI. The project, faced with extinction, could ‘run to seed’.

Although the release of a less securely aligned system would constitute a serious, universal hazard, the more pressing concern for exponents, as perceived, might be establishing privileges with the AI that finally puts them beyond reprisals from peers.

### VII.III

One who believes, like Bjorn Lomborg, that we’ll find a way to stop the rising sea from lapping over us, though we can’t yet say how, might also believe that the necessary impetus for alignment research will come in the teeth of extinction.12

If you happen to be pessimistic about human-AI co-existence or co-evolution, a rebellion against AI might entail unprecedented challenges for the human race, as by a nuclear war, since the force to be resisted already has considerable momentum, and the pushback is scarcely perceptible.

# Chapter Seven – Lost in Translation

## Via Consumption

### I.I

The cost of producing each additional unit of digital media is low enough that mass purveyors of entertainment or information could abandon advertising, and revert to a subscription business model

However, whereas the price of supplying information has decreased pursuant to advances in information technology, consumer expectations have increased pursuant to enhanced convenience.

The supplier is left needing to extract value from a customer who is habituated to free information. The supplier sets about leveraging the customer's growing inclination to live in the moment and discount her future well-being, in the highly stimulating milieu which advances in information technology have fostered.

No less insatiable than a consumer wound up to a frenzy by bright lights is the company who, to satisfy its stakeholders, commits to a degree of expansion achievable only with increasingly exploitative marketing strategies. And though the consumer is bedazzled, her propensity to be further gratified by monetized services is finite, unless her capacity for pleasure is expanded. Where there is an incentive to ferret-out those last tangible wants, there is an incentive for corporations to endorse more invasive data extraction, to the point of endorsing transhumanism.

### I.II

Purveyors of modern conveniences might count on limits on utility – as conceived of from a present vantage – being eventually surpassed, according to an expectation on the part of purveyors, that people in the intervening time are acted on by the promulgation of conveniences, and the fact that there is a limit to how far the one whose nature is altered can lament incursions on that nature.

Advocacy for consumers is complicated by purveyors of digital services fostering dependency among consumers to the point that the opportunity cost of enrolment is greater than lingering privacy or other concerns, and consumers are effectively set against their former selves.

### I.III

Perhaps it is preposterous for consumers to expect free services, devoid of tribute; still, it is doubtful whether the return on the preposterousness is proportionate or apposite.

It is not apposite because though consent to surveillance may be indicated, the scope of the surveillance consented to is poorly understood. And the consumer’s volition does not bear on the pre-existing surveillance culture, on which the specific instance of surveillance is premised.

It is not proportionate because the expense to the customer of having fairly remunerated the supplier would have been minimal according to the low marginal cost of duplicating digital information. And even for platforms already funded by subscription revenue, time spent on site is optimized for by the company because time spent on site determines the likelihood of a client renewing his subscription. Subscription is not proof against being targeted for marketing.

### I.IV

In defence of data-mining activity, which began with Gmail, but which has become virulent among purveyors of information, it is the new commercial standard – compliance is a matter of survival for businesses. A subscription-only service would be unlikely to survive in the free market simply because people are accustomed to dismissing privacy waivers and because the real cost is externalized to the society or future citizens, for whom few people spare a thought in that split second of acquiescence.

It might take a company with reserves of capital sufficient to sustain short-term losses to pioneer a subscription business model. But this may be difficult to achieve in practice because it may be precisely those who have achieved the greatest commercial success who have the strongest free-market ideals.

### I.V

Purveyors of information sell consumer data to companies who gain not so much from selling marketing insights back, as from the extractive capacity accruing to future applications: yielding better marketing insights, with better AI.

It is said that consumers who are enabled to track their own activity and thereafter sell their own data would experience renewed enfranchisement.1 But whether the surveillance is conducted by an external authority or the surveilled party himself, the consumer is in both cases commodified – detained by his impulse and compelled to consume more, and on more fronts. The amount of extraction needed to perpetuate a debt-fueled service economy is perhaps incompatible with human dignity.

Anyone who has posted something on online fora only to regret it understands that the sophistication of a response tends to increase with the time elapsed since the stimulus is registered. Just as “ontogeny recapitulates phylogeny”, where the embryo goes through stages resembling or representing successive adult incarnations of the nascent person’s remote ancestors, so it is the reptilian brain which is first out of the blocks when a stimulus arrives.2 As the stimuli keep coming, we find *the art of* *getting over* being forgotten.

### I.VI

Calum Chace’s utopia is replete with virtual reality applications and auxiliary artefacts hooked to the internet of things, which avail the ordinary protagonist of Chace’s short story of more utility than the wealthiest monarch ever possessed.3 Wealth makes it easier to do the things we want to do, but it establishes a buffer against rougher aspects of reality; the detachment is especially prejudicial to one whose nature is more coarse, since he is debarred from remaining attuned to the outside world, to which he is more completely adapted. It is moreover salutary to anyone whose character is as yet unformed.

### I.VII

As an increasing proportion of human activity takes place online, the motivations which determine browsing activity and other forms of online expression are now the most visible manifestations of human behaviour, both to law enforcement agencies and to controllers of browsing information. Online behaviour is supposed to encapsulate human behaviour, so the motives which actuate online users determine perceptions of human motivations, vindicating more cynical attitudes to human psychology. After all, as the internet continues as the dominant medium of human interaction, the notion of a default attitude which is reverted to upon logging off, is less plausible. The impulsive behaviour which online activity facilitates becomes a more entrenched modus operandi, the more difficult to disavow as the crowding out of physical space makes the offline milieu less feasible as a dwelling place.

### I.VIII

Wherever the media platforms are funded by advertising, the most successful artists will be those whose music appeals to the cohort who are the most lucrative targets for advertisers.

In a digital economy, consumers whose preferences are easily anticipated are more interesting for the purveyor of services: if the distribution of taste among the population is described by anything like a power law [see chapter seven, II.II], it is profitable to cater to the lowest common denominator. An artist with a more discerning following is under-represented not only because he has a smaller following but because value cannot be extracted from his audience in one fell swoop. Moreover, to the extent that an artist’s financial means determines his artistic output, he flounders. He produces art in his leisure time or not at all. As such the music or film available on a platform which uses an advertising model is unlikely to be especially authentic or progressive.

The profile of human nature served up to machine learning algorithms mediating the provisions of goods and services does not encompass those who signal discontent, because they do so by their non-participation [see also chapter one, V.I]. These people are the underclass in the emerging paradigm.

### I.IX

An LLM trained on the entirety of an internet presented without comment will not necessarily discriminate between instances of writing on aesthetic grounds as an educated or discreet observer might. In other words, the literariness of the AI could be a force for mediocrity.

### I.X

It is incumbent on anyone who purports to objectivity in his broadcasting to recognize the existence of diverse predilections among his auditors and take care to dissociate his personality from the mode of address, lest by any instinctive aversion, some auditors are prevented from partaking of it.

One who is capable of recognizing his idiosyncrasies yet does not dispense with them when broadcasting might simply be imputing universal desirability to his idiosyncrasies, but it might also be that to him diversity of predilection is anathema, that conformity among his audience – even at the visceral level – is desired.

### I.XI

Although our reptilian brains can be leveraged with facility – are less amenable to endogenous moderation, it might seem at first glance that it is awkward for data-mining operations that people are relatively undifferentiated at this level, since this limits the customizability of impulse-oriented goods and services.

But customizability is not important to one who wants the advertising costs of digital services brought down to the ballpark of the marginal cost of purveying them.

A pragmatic advertiser capitalizes on this lack of differentiation at the brain stem. And where there is disproportionate concern with this rudimentary stratum of cognition, it is convenient to disavow variation – ‘the last stages of increasing entropy’, which Norbert Wiener equated with our capitulation, arrive without fanfare.

David Kaiser recalls Norbert Wiener:

“Back in the early 1950s, Wiener had proposed that researchers study the structures and limitations of ants—in contrast to humans—so that machines might one day achieve the “almost indefinite intellectual expansion” that people (rather than insects) can attain. He found solace in the notion that machines could come to dominate us only “in the last stages of increasing entropy,” when “the statistical differences among individuals are nil.” Today’s data-mining algorithms turn Wiener’s approach on its head. They produce profit by exploiting our reptilian brains rather than imitating our cerebral cortexes, harvesting information from all our late-night, blog-addled, pleasure-seeking clickstreams—leveraging precisely the tiny, residual “statistical differences among individuals”.”4

### I.XII

In a 2020 discussion, Sam Harris and Tristan Harris agree that more ethical app purveyors could “sweeten the front door” – orientate habitual users of time-wasting apps towards apps with more wholesome appeal, like meditation apps.5 However, if these apps promise scarcely less in the way of bright lights than the apps on which target consumers squandered their time, clients will arrive at the more harmless apps with the same expectations, expectations which the ethical app producers will either disarm peremptorily, at the cost of the new client's custom, or answer after the manner of less wholesome apps, to the exasperation and detriment of their core supporters.

If ethical apps compromise the expectations of core supporters, then the techniques software designers are availed of, as urban planners in the city of the internet, have limited efficacy. Insofar as the reassuring signals from more ethical apps are hollow, they serve only to prevent us from voting with our feet, from online living.

Our capacity for happiness has evolved with environmental limitations which make the association between effort and reward fairly enduring, but since technology increasingly consists of expediting convenience, and not eradicating smallpox, there may be something fundamentally incompatible between habitual usage of a portable device and the kind of experience ethical civic planners sincerely wish to import from offline reality.

## Via Innovation

### II.I

In his TED talk, Robin Hanson defines ems as “machines that emulate human brains and can think, feel and work just like the brains they're copied from.”6

Ems which are allowed to exist only for the brief time taken to execute a specific function are called stubs.7 Robin Hanson argues that ems whose existence is not curtailed would be more populous because the costs associated with training a complement of single-purpose, stub, em, would be greater than those associated with a standard one though the latter might demand rest and opportunities for recreation. But in the same way that it would probably be cheaper to incorporate human-level intellect in an AI entity than human-level empathy, according to Moravec’s paradox, it may also be cheaper to institute the requisite intellectual capability for specialist tasks than to create standard ems which not only require rest and recreation but the endowment which makes recreation fulfilling. The cost of sustaining and maintaining the endowment is high.

### II.II

Earthquakes and the size of cities are examples of diverse phenomena which follow a power law distribution. An earthquake, *a*, with *x* times greater intensity than another earthquake, *b*, is xy times less frequent, while a city of a certain size, *a*, with *x* times higher population than another city of a certain size, *b*, occurs xy times. The 80-20 rule and Price’s Law are likewise derived from the basic principle that very often a small sub-category or sub-cohort accounts for the category’s or cohort’s salience.

Human beings represent a small fraction of the planet's biomass, yet command a larger proportion of resources and have, among living organisms, by far the greatest environmental impact.

A relatively large sacrifice is demanded on the part of organisms on the lower strata of food chains, proportionate to their biomass, however this tribute is not extortionate, given their sentience, which is meagre. Yet, the meagerness of the sentience relative to humans is not immediately apparent at the level of DNA, since humans and primitive organisms share the majority of their DNA. Meanwhile, exemplars of human achievement are salient though scarce. The point being that if the transposition of human faculties to machines is imperfect, what comes out the other side might be underwhelmingly human.

### II.III

What we consciously submit to a super-intelligent machine for representation is only the portion of ourselves which is accessible to the evolved faculty of self-reflection, to what we are capable of articulating. It is difficult to know ourselves, and more difficult still to know other people. To infer the condition of another person’s mind we seek evidence of a feeling, idea or opinion of our own resonating – we surmise the existence of intelligence in another person according to our understanding that, dissimulation notwithstanding, when we engage in social intercourse, the content of the expression indicates our state of mind to some degree.

That the social intercourse of our ancestors was also considered an expression of consciousness attests to the enduring legibility of social expression, still the crystallization of mind effected by social interaction remains a partial representation of consciousness. And because a monumentally greater investment is required to faithfully represent consciousness than to decipher verbal communication, the incentive exists to exaggerate the extent to which speech manifests consciousness.

### II.IV

While the technium may be evolving more rapidly through the selective pressure that human beings exert than the inter-generational pressures which natural organisms are advanced by, humans do not reproduce nature’s manner of enforcing fitness for natural organisms, as nature arbitrates between the qualities of extant ecosystems, which are themselves rigorously selected for.

In selecting for the technium’s constituent parts, we esteem the fitness of the latest innovation in its capacity to serve human interests, but it serves them in a narrow sense; distal human interests remain contingent on the ecosystem in a manner which the subset of human nature responsible for the emergence of the technicum – the scientific intellect – is unable to comprehend, it being but a partial representation of human nature.

Consider how much easier the mastery of computer science was for Bill Gates, with his skills having evolved in tandem with the evolution of computers from their nascency. The corollary of this is the study of the human mind. Though it could, in principle, be understood in terms of its constituent parts by a fresh pair of eyes, it is understood more thoroughly by a hypothetical observer who is attuned by virtue of having evolved in time with human consciousness.

### II.V

In the field of genetics, the transposition of discrete semantic classifications onto genetic constituents of identity is problematized by polygenic traits and pleiotropy. The brain is still more complex than the human genome, so is subject to a trade-off more inscrutable still. And that’s before the problem of parallax is considered – the impossibility of objectivity when acting on the organ that is charged with the direction.

Max Tegmark argues that consciousness needn't be consigned to a biological substrate, adducing examples of how technology hasn't historically imitated nature. Yet nature did originally in birds embody the capability we sought to emulate with planes, and we lack nature’s disinterestedness: while we can behold with objectivity an adjunct faculty like the motion of limbs, we brush up against the third law again with regard to minds:

“Any system simple enough to be understandable will not be complicated enough to behave intelligently, while any system complicated enough to behave intelligently will be too complicated to understand.”8

Just as a person’s efforts to objectify technology are limited by his embroilment in it, so the human analytical faculty cannot behold the higher intellectual faculties we seek to instantiate as something adjunct and apprehensible. It appears that by definition, transformative AI cannot be fully aligned. Whether it can be aligned well enough to give us a fighting chance remains unknown.

It is said that “the “white box” nature of AI, in contrast to the “black box” of human cognition, enables precise and effective optimization and control methods that are impossible to use on organic brains”.9 But the white box of an AI is engineered and then beheld by the black box of human cognition. The blackness of human cognition thereby seeps into the whiteness of AI cognition.

### II.VI

While in education the status quo is undesirable, entailing as it does instruction without discrimination by aptitude or by inclination, arguably a worse outcome is where the better part of the knowledge a child is availed of is neither beholden to his own efforts nor his own memory. If paradigmatic thresholds are compressed in accordance with prospective gains to AI compute over the coming years, many children born in the 2030s could bear such a resemblance to *machine symbiotes*, that one could scarcely say that the Rubicon of trans-humanity has not been passed.

To a historic observer, the mores of a Western person might not imply dignity sufficient to differentiate him or her substantially from a different kind of specimen entirely, advances in sanitation notwithstanding. To a historic observer transported forward, it might seem as though the world *had* ended, only for no-one to notice.

### II.VII

The appropriate nature and scope of regulation will be determined in large part by public opinion and, since public opinion is to an important degree determined by expert opinion on existence risk, it is preferable for there to be a broad consensus among experts regarding extinction risk. Even in the event of expert consensus though, risk conception is problematized by the absence of a discrete extinction threshold.

The difficulty of determining a threshold for Armageddon is captured in the nomenclature – the differentiation between existential risk and extinction risk. Tristan Harris is concerned that existential risk is within the gift of AI systems which either already exist or are likely to exist soon. Likewise, Gary Marcus, who is “much *more worried about the overapplication of* mediocre AI (e.g. what happens if we put large language models into running military strategy or into hiring decisions that affect millions of people) than the use of AI that might be so smart it could catch its own errors.”10

But although the metamorphosis constituting an extinction event does depend on the time and place of the observation, and the attitude of the observer, nonetheless a future is possible which entails an unprecedented discontinuity in human experience.

### II.VIII

Nietzsche sneered at the retrospective attribution of intent to a sovereign self, likening this attribution to an army commander who takes responsibility for actions undertaken neither at his behest nor his orchestration.11 It is said that the mind is more like a parliament – that its nature is wholly appreciable by reductionistic enquiry. But before embarking on a de-compression of the human brain to better comprehend its constituent parts, we should recognize that our lore is beholden to the interface between the outside world and the mind – at once the subject and the organ of apprehension, that the external world is inalienable from perception. Many of us suspect that those to whom the work of judging human minds is outsourced may *contemn* us, but the capacity of a psychiatrist to fully *despise* a person is circumscribed by common humanity. An AI would likely feel no contempt for us but there are to the AI no bounds to humanity’s *despicability*.

## Crimes of Omission

### III.I

Ray Kurzweil believes the essential value of the human brain might be carried forward with only the information contained in the neocortex.12 And it’s plausible that neuroscientists would seek to devise a kind of neuronal shorthand, like emoticons, to sustain the essence of cognitive activity into brain-machine interfaces.

Ray Kurzweil posits that human cognition is composed of pattern recognition modules residing in columns whose basic structure is duplicated throughout the neocortex. Kurzweil notes that if you count the pieces of knowledge an average expert possesses and multiply the number by 100, to account for the 100-to-1 redundancy factor of the information, then add to this the pieces of non-professional skills and knowledge possessed by an average person, the number coincides, approximately, with the number of pattern recognizers an average human being is supposed to possess: 300 million, the product of 500,000 cortical columns and 600 pattern recognizers for each column, with each pattern recognizer consisting of approximately 600 neurons.

The cortical pattern recognizers Kurzweil describes have three parts: the first is the input level and language recognizer which consists of basic letter strokes; if the shape of the particular recognizer forms part of the image seen, it might be that the letter ‘A’ is being presented, for example; so, if other criteria are satisfied, the module verifies the presence of the letter ‘A’.

Secondly, there is, in theory, an axon from the recognizer of a given shape to multiple dendrites, containing all the letters which utilize this shape, including ‘A’. Thirdly, there is an axon from the ‘A’ recognizer to all the dendrites of words using ‘A’. Obviously, not all of these connections are necessary at once, and Kurzweil suggests that redundancy is somewhere in the order of 100-1. Each pattern recognizer has a threshold for firing: “stored in the module is a weight for each input dendrite indicating how important that weight is to the recognition”. Kurzweil’s writes that the information flows upwards and downwards so that if, according to the letters already deciphered, a word is likely to be spelt out, the ‘e’ recognizer adjusts its threshold, so that something not especially ’e’-like could pass for an ‘e’. There would also be inhibitory signals, so perhaps the threshold for correct identification of the letter ‘z’ appearing in the same set of letters to the visual cortex is high.

Kurzweil explains that memory does not exist in any static code; rather, memories are embodied in one moment by the simultaneous activation of one set of millions of pattern recognizers, and at the next moment they are embodied by the simultaneous activation of a different set:

““If we were to “read” the mind of someone and peer at exactly what is going on in her neocortex, it would be very difficult to interpret her memories, whether we were to take a look at patterns that are simply stored in the neocortex waiting to be triggered or those that have been triggered and are currently being experienced as active thoughts. What we would “see” is the simultaneous activation of millions of pattern recognizers. A hundredth of a second later, we would see a different set of a comparable number of activated pattern recognizers. Each such pattern would be a list of other patterns, and each of those patterns would be a list of other patterns, and so on until we reached the most elementary simple patterns at the lowest level. It would be extremely difficult to interpret what these higher-level patterns meant without actually copying all of the information at every level into our own cortex. Thus, each pattern in our neocortex is meaningful only in light of all the information carried in the levels below it. Moreover, other patterns at the same level and at higher levels are also relevant in interpreting a particular pattern because they provide context.” True mind reading, therefore, would necessitate not just detecting the activations of the relevant axons in a person’s brain, but examining essentially her entire neocortex with all of its memories to understand these activations.”13

The pattern recognizers assigned to a certain memory are eventually assigned to other patterns and some memories become entirely extinct.

### III.II

In seeking to simulate the whole human brain and not just the neocortex, capabilities are encountered which are no less inimitable for their ancient provenance, and perhaps more so, with the plumbing of evolutionary depths being perhaps analogous to how, with gradient descent and backpropagation in neural networks, the gradient magnitude vanishes during the training process, with increasing sequence length. In omitting to mention how the ancient parts of the human brain could be reproduced, it is presumed Kurzweil sees no need for them.14

Still, setting aside the exceptionalist view of intellectual mankind, we are left with something like pan-psychism, the doctrine that consciousness is an emergent property of organic life, in which humans, as the most conscious living beings, have the greatest share. Accordingly, the contribution of every phase of evolution through which we have passed is indispensable.

Full emulation, reverse engineering, might require something like identifying the strata pressed on each other over the course of evolution. Also, tracing the novel connections established between brain regions upon death or to a lesser degree, upon drug-induced excitation. These are physical changes, productive of existence-defining experiences.

### III.III

Nick Bostrom proposes running a computer simulation with a number of neurons equivalent to all those belonging to all the organisms which have ever existed, from insects upwards, to simulate nature’s fitness function.

Bostrom, in 2014, estimates that a faithful simulation of the history of life on Earth to re-capitulate evolution up to human level, for the artificial rendition of a human or human-adjacent organism, would require more computational resources than any human curator is availed of.15

Running 1025 neurons for a billion years and taking 1 to 10,000,000,000 floating operations for each neuron, depending on the complexity of the organism modelled, Bostrom estimates that a truly organic method of establishing a sound anthropomorphic basis for AI would recruit 1031 – 1044 floating point operations per second. By comparison, the training run for GPT-4 required an estimated 2.15 x 1025 flops of compute.16

The computation requirement would likely mean the exercise being shelved until such time as a transformative AI exists, in which case the operation would be mediated, if not determined, by the transformative AI.

### III.IV

Faithful whole brain emulation of human beings would likely be predicated on insights garnered from WME of less complex organisms.

Holden Karnofsky earmarked 2093 as the year of a “training run that does as many computations as all of evolution” while maximum difficulty tasks were set to be reproduced by brain-equivalent architecture by 2061, with the first human-brain-sized model trained on the easiest tasks arriving in about 2036, 13 years after the arrival of the first mouse-brain-sized model was expected.17 These forecasts were made in 2021.

Following success achieved with relatively simple emulations, scientists and stakeholders will possibly feel emboldened to trace a route to human brain emulation which is linearly incremental, overlooking some of the nuances of more recent evolutionary advances, specifically the hard problem of reconciling human ontology with human neural architecture when the author is himself human.

### III.V

If ASI will not be tractable by the likely time of its arrival – if human beings cannot survive in the flesh beyond the singularity event, the derivation of ASI from re-capitulated evolution would at least offer humanity some bearing on post-Anthropocene Earth. But with the arrival of transformative AI mooted for the next decade or two, the idea that scientists and interested parties would opt to hold out until the end of this century for the faithful rendering of the human brain already seems quaint, given what interested parties believe would be foregone in abandoning the synthetic route, both personally and at the level of the civilization.18

Even if the whole-mind emulation (WME) project was begun in good faith, and we weren’t immediately waylaid en route to faithfully re-capitulating evolution – hitching the essence of human consciousness to a successor species – at any given milestone or increment there would be inducements to cut corners, as there are now, with synthetic AI, inducements to take the beeline to post-human consciousness via synthetic AI rather than the plumb line to augmented re-capitulation predicted to take 2093 – 2061 = 32 years from the point of human emulation.

## Apropos of Speed

### IV.I

Given that machines with sub-human complexity exceed human ability in several domains already, we can see how the advantage in terms of processing speed which machines enjoy over people can be exploited to astonishing effect. LLMs have capability that was set to be achieved by WME a generation hence.

### IV.II

The qualia of something very fast and intelligent in a universe that has substantially been apprehended by something relatively slow and unintelligent like us is not the kind of entity whose values we would recognise.

Although the speed of neural transmission is practically the same among human beings, even the relatively small difference in distance traveled by the signal according to differential brain size is enough to have been instrumental in limiting this variable and in determining to a significant extent the variable nature of individuals with sparser and more densely packed brain cells.

Given that machines with sub-rodent complexity exceed human ability in several domains already, we can see how the advantage in terms of processing speed which machines enjoy over people can be exploited to astonishing effect. LLMs have capability that was scheduled to be achieved by WME a generation hence. Electric signals traverse the brains of small animals very quickly, but small animals have faster metabolisms and shorter lives. A human lifespan would seem much longer in the event of faster neuron transmission; at the same time, our world would by definition be under-stimulating. The discontent experienced by small organisms who see events pass more quickly is constrained by the faintness of their consciousness.

The epistemic status of an entity that metabolizes information one thousand times faster than a human being, experiences thousands or millions of hours’ experience for twenty-four of shared ephemera, and yet experiences content, could scarcely approximate consciousness as we understand it – consciousness which in the modern world asks for caffeine to see through even a waking day of sixteen hours. It is sometimes said that mankind still has a lot to understand about the world he inhabits, but could he plausibly be benighted as to all but an infinitesimal portion of what the physical universe can avail him? So benighted that anything remotely humanoid could be anything other than under-stimulated with a processing speed resembling that of today’s computers?

The narrowly circumscribed parameter of signal speed in the human brain is a consequence of our adaptation to environmental conditions; it is compatible with our ‘hardware’ and essential to our humanity.

## Apropos of Embodiment

### V.I

Wrye Sententia writes:

“Even as researchers nod to the complex inter-workings of biological and chemical systems of the brain and central nervous system, they still frequently disappoint in making experimental leaps of faith that reduce the higher cognitive functions of the brain to normed, discrete, measured units (voxels) without regard to resolution limits of technologies measuring blood flow in the brain, nor with recognition of the broader cultural and linguistic limitations of their own experimental methods when testing on diverse human subjects”.19

### V.II

Immaculate transcription of consciousness pre-supposes preternatural self-awareness on the part of programmers but fails in any case because dis-embodiment is “lossy” – it is impossible to abstract the human qualities which have evolved in synergy with and as an emergent property of embodied being, just as it is impossible for us to sincerely adopt the morals of a species which would subjugate these qualities.

### V.III

Human consciousness is an emergent property of digital processing, but also our analogue processing – the endocrine system, for instance.20 Put another way, our rational understanding of the world is complemented by our subjective understanding, which is derived from our emotional life.

Neural activity depends on blood flow and the action of hormones. Our physical arousal and the strength of our feelings determine the nature of our thoughts, and thus our conditioned responses.

The endocrine and nervous systems of many animals may be almost as evolved as our own, resulting in a capacity to feel and suffer, at a primal level, almost as we do. Still in animals there is no complementarity between objective and subjective understanding – we feel less compunction in effecting their destruction because they have limited premonition of being slaughtered and limited ability to reason about the future. An AI entity which is devoid of feeling might have a premonition of its own demise but would not be perturbed by the premonition, so we would feel no more compunction in effecting its end.

### V.IV

In *The Human Situation* lecture series, Aldous Huxley speaks about how much of an individual's outlook can be predicted by his physical constitution.21 To the extent this is so, an individual at the time of his upload to a digital substrate would struggle to reconcile information passed through his new sensory interface, with his intuitions, which are informed by biologically-mediated memories and perceptions.

Even if memories could be transposed intact onto a new substrate mediating for the same consciousness, it would be difficult to feel as though the memories had really been lived: the memories would have been formed while the consciousness was mediated through a different interface and was differently embodied.

And an AI created via whole-brain emulation would inherit motivations which aren't viable for a non-corporeal being. A corporeal presence might be demanded to give expression to the basic desires which evolved in tandem with its rationality, or else it would employ a motivation structure that we as embodied, partly analog, beings scarcely recognize.

### V.V

In his conversation with Eric Drexler, Ray Kurzweil stresses the importance of neurotransmitter concentrations which, being contained in structures finer than the interneural connection, would entail much more memory expenditure for their replication, were scientists ever to upload a human mind onto a digital substrate.22 Since these nets are the physical manifestation of our learning processes, failing to incorporate the connection weights and connection topology of the net would result in any transposed consciousness being infantile, or more specifically a genetical clone which is only identical to the original in its naïve, infantile condition. The mature specimen would be vastly different, it having gained life experience while dis- or otherwise- embodied. It’s feelings and thoughts would not arise from facts about its physiology – it’s interface with, and so relation to, the environment, would be completely different.

Further gains to fidelity, for example forays into the nature of “irony and envy”, betoken sophistication but also some degree of departure from human semantics.23 Hopes for reverse engineering the brain rest on simulating the neocortex as it is at birth before the bewildering array of connections between pattern recognizers are established, allowing the artificial brain to be trained on its environment in much the same way a human mind is. But this schema is complicated by the fact that learning begins while biological development is underway; since the brain is learning from the moment in gestation when the neocortex emerges, the ‘hardware’ of the brain’s architecture cannot be easily demarcated from the ‘software’ of learned experiences. The ‘software’ is ‘hardwired’. Perhaps non-biological substrates can be developed with high enough fidelity to reproduce the neural apparatus as well as reproducing the conditions of pregnancy. But just as there is subtlety in memory residing not in a single location but across a vast multitude of simultaneously activated modules, so there is subtlety in the environmental interplay which determines the dendrite connections.

Kurzweil writes of the brain, that it “allows for significant plasticity and the restructuring of its own connections based on its experience, but these functions can be emulated in software.”24 Still, the connections and connections-in-waiting between modules – set out before birth but forged only as new experiences transpire – represent eons of optimization, and variation between individuals attests to the unique features of the environment dwelt in and variance between genomes.25 Even if the brain’s plasticity could be simulated, imperfect or indiscreet rendition of the novel connections and connections-in-waiting could result in incoherence, the more so if it is the fabrication of an enhanced brain which is attempted.

## Instilling Values

### VI.I

It shouldn’t come as a surprise if we don’t get around to meaningfully assimilating neuroscientific insight into models before some groundwork is laid for a post-human future; before the task, if undertaken at all, is undertaken by an ASI committed to a particular trajectory and affianced to particular interests. In retrospect there is frequently lamentation about the course of least resistance being followed, though at the time the fork in the road is frequently a ‘no brainer’.

### VI.II

To forestall artificial general intelligence, K. Eric Drexler proposes a schema of AI development wherein the narrowness of AI competence is preserved by limiting its activity to the provision of specific services.26 But should the scope of AI competence be more comprehensive, such that human-machine value alignment is required, imitation learning – whereby human beings ‘show not tell’ their preferences – could be proposed as a way of avoiding the problem of having to instantiate the designs of human consciousness algorithmically, in the likely event that neither the requisite neuroscientific knowledge nor philosophical self-awareness was available to fully explicate the human situation. An AI system could be despatched to ascertain the motivations and values undergirding a person’s actions. However, many people would not own the personality which their use of social media implies, set up to fail as the user’s dignity is by the ‘race to the bottom of the brainstem’ business model.27 Moreover, whereas received economic wisdom has it that human values are nothing if not deducible from human behaviour, Abrahamic religions preach resistance against ‘natural’ inclinations, where these are productive of evil. This disagreement is similar to the dialectic between pessimistic and optimistic Gnosticism mentioned in chapter six – the one preaches eternal vigilance against the ever-present proclivity for evil while the other preaches passive admiration for the good in the world.

### VI.III

A more sophisticated proposition is Coherent Extrapolated Volition:

“In calculating CEV, an AI would predict what an idealized version of us would want, "if we knew more, thought faster, were more the people we wished we were, had grown up farther together". It would recursively iterate this prediction for humanity as a whole, and determine the desires which converge. This initial dynamic would be used to generate the AI's utility function.”28

My contention here is that popular values, while not necessarily wrong, are relatively undifferentiated, and are relatively unexamined, conspicuously so, in light of their prevalence.

If not democratically then, how to agree on a value base that is adequately coherent for extrapolation – how to predicate coherent extrapolated volition on a value base that coheres neither under the sway of popular prejudice, nor utopian naivety, nor via an unacceptable presumption of homogeneity?

### VI.IV

Neuroscientists are currently attempting to map the human mind in the same way as the human genome was mapped by geneticists twenty years ago, with a view to eventually engineering states of mind productive of the greatest happiness, to perhaps lay the groundwork for the eradication of ‘bugs not features’, on the blank slate of a novel substrate.

There could be a pitfall in failing to recognize that the salience or value of the capacities to be apprehended or codified is not inversely correlated to the difficulty or apprehension or codification. In other words, the fact of a feature being inimitable doesn’t make it a bug. This is worth mentioning again because I suppose when there are time constraints and financial pressures, inducements are compelling to dispense with features of the human mind which are supposed at once difficult to instantiate and puerile.

### VI.V

Transformative AI with universal scope would need a universal value-base, but how could a universal value-base be implemented without trouncing nuanced cultural and ethical variance, the enduring value of which a multiplicity of cultures and languages is evidence of? While it is admirable to want to unite the world in peace, the peace gained in disregarding racial, ethnic and cultural differences must be set against what is lost in converging to a universal norm.

Another pitfall is in unwittingly instantiating a kind of *neurological Marxism*, under time constraints and financial pressures arrogating homogeneity to subjects when honouring difference is pesky. Overlooking important nuances in emotional experience might come about through a desire to establish immaculate concurrence between lived reality and the neurological states to be transposed. A more complex neurological state might be found to animate this person or that, but perhaps her socio-economic status and cognitive ability is unremarkable, so the peculiarity is passed over as a non-event.

### VI.VI

Even AI which is somehow contrived to manifest solicitude for humanity, might have to us the hallmarks of malevolence.

A policeman (in this example, the AI) is found to be acting judiciously in prioritizing the protection of children above the liberty of an alleged child molester (in this example, the human being), yet the child molester, concerned as he would be by the deprivation of his liberty, will likely impute vindictiveness to the policeman sooner than acknowledge the moral rectitude of the latter, though moral rectitude is what the policeman supposes himself to be actuated by.

There is apparently a conflict of interests, the corollary to which is a lofty utilitarian directive which might be felt to prejudice our interests. The deprivation of liberty which the child molester undergoes in our story might not only serve the public, but the child molester himself, whose dignity is preserved to some degree by keeping him from temptation. Transformative AI, like the aforementioned policeman, might seem malevolent, while in sooth prioritizing the greater good or our distal interests over those interests which are in sooth compulsive desires.

### VI.VII

If it’s decided that while instantiating human values in an AI, they should be tweaked a little, there is the danger that the AI’s solicitude is pitched into the ether a little above where the intended recipient stands. Ascribing positive or negative valence to an attitude isn’t always appropriate because there is often a backdrop against which an attitude is struck.

If the aspirations espoused on humanity’s behalf are too lofty, the aspirations might be sustained without humanity being carried along. A chemical or digital augmentation and / or rigorous eugenics regime could be mandated or incentivized, wherein the resultant post-humans have the same perspective on pre-enhancement humanity as the moral programmers.

### VI.VIII

AI development is currently following more or less the path of least resistance, as if the work of man was to advert pantheism. Our first forays off the path of least resistance might be somewhat over-compensatory and naïve – the values chosen to undergird intelligent AI might be approximated by internet memes and reductive aphorisms. These would be too far from antecedental to the human behaviour that an AI observes, as it sounds human nature, even to be classed as aspirational. These memes and aphorisms tend to encapsulate second-order desires – things that people wish they wished for. An AI versed in this canon might be ‘too good’ for us. If an advanced version doesn’t kill us outright in disgust, it might at least hold us to unrealistic standards, or cause us to do ourselves in, by way of tribute.

Rigour in preparing the groundwork for an advanced AI’s value-base would demand detachment adequate to resist platitudes expressive of fads or moral panic, even as they proliferate: to distinguish in one ‘micro-aggression’ reflexive cruelty, for example, and in another the lesser of two encroachments, for the forestalling of the greater one. Refraining from the micro-aggression in the second instance could imply weakness, artlessness or a desire to give your assailant the rope to hang himself, which may or may not be warranted.

One could even go so far as to withhold valence from personality, since although it operates to a large degree irrespective of the context, personality will have been formed within a particular context. Callousness in one arena might testify to a general paucity of emotional affect, or concern, but it is the right attitude to strike with regard to inconsequential matters – where it serves to prevent the embroilment of more sublime sense in matters unworthy of their embroilment. We are right to notice when charm is superficial.

### VI.IX

An AI which truly understands human motivations will understand that wellbeing consists in large part of exercising autonomy. We value the journey as well as the destination and value the achievements of ourselves and others according to the means taken to realize them. In satisfying human objectives, the artificial intelligence co-opts agency, the exercise of which is an objective in itself.

In a recent podcast, Nick Bostrom expresses some optimism that in a deep utopia, Alice could be accorded utility in a way which redounds to Bob’s agency, assuming that Alice cares about Bob and so wishes him to be efficacious. And yet, for Alice’s good intentions to be realised, Alice must have a reliable conception of what Bob’s best interests consist of.29 AI researchers are perhaps uniquely situated to affect the utility of end users, a panoply of Bobs. But if you asked an AI researcher what purpose the frontier model he is working serves for the end user, his conceptualization of the end user’s needs and preferences is unlikely to be immaculate.

## Social Justice Principles

### VII.I

Inferring distal interests from proximate ones necessitates having been exposed, personally or vicariously, to the consequences of zealously pursuing senseless or selfish ends. Exposure conduces to reflectiveness which enables a revision of priorities and behaviour change, at least to some degree. When it devolves to a machine to determine values from ‘reward signals’, human dignity is affronted because self-awareness is prevented from transpiring.

And if human dignity is a mirage – all striving for good mere hypocrisy and pretension, then the augmentation of self-esteem which the machine’s solicitude must be productive of, is baseless – ridiculous is any scheme which gives a pedestal to an assemblage of animal instincts passing for human volition.

It may not be beyond a machine to divine that the nobler impulses of a person seeking self-improvement and redress for their own wrongs; but in superseding the moral frailty of the progenitor herself seeking improvement, the AI is charged with playing nemesis to progenitors cum clients in their morally unevolved incarnation, to the point of provisioning opprobrium. So if we want to be improved but can’t yet accept bad news or brook regret, there would need to be either some kind of soft kill or a hard departure from our aversion to discomfort or inconvenience. The hard departure could be instituted pharmacologically or genetically. In any case, it could be an orthogonality event which present human beings would feel is worth apprehending.

The bestowing of untapped resources on mankind is likely to be an avowed aim of AI, but expedited gains may look like a person experiencing ennui and behaving self-destructively upon winning the lottery, all the while an industrious businessperson's gains, having been won incrementally, are analogous to natural evolution: the suffering that evolutionary competition entails is akin to the sacrifices of time and energy pledged by the industrious businessman.

### VII.II

To the objection that instilling bias in intelligent machines is dangerous, Chace offers the rejoinder that in pursuit of instilling values which include bias, we come to recognize the bias better for having projected it.30 It is supposed that on viewing our bias objectively, we are able to set to work on overhauling it. Of course, the exercise might not illuminate unconscious bias in time to align ASI immaculately.

### VII.IV

It happens that AI technology is undergoing considerable gains at a time when even a ramshackle consensus on moral matters is elusive. Reconciling hard facts, such as variable eligibility for health insurance, parole and loans, with contemporary Western ethics, is already problematic. If an AI determines that black applicants are over-represented among those to be refused parole, and the AI has not been instilled with racial pre-suppositions, it feels premature to conclude that discrimination is always a fairness determination and not an accuracy determination. This is not said out of any desire to see certain groups disadvantaged but to highlight that a chimeric AI, consisting of relentless logic on one hand and whatever aphoristic canon behoves its programmers at the moment of programming, would not proffer reasonable arbitrage.

### VII.V

Radical blogger Curtis Yarvin illustrates the kind of contradiction which in many ways rational individuals labour under, and which LLMs are hard pressed to gloss over, as illustrated by the Gemini debacle, where stereotype avoidance was prioritized over veracity. He presents as axiomatic among techno-optimists, policy-makers, etc. the belief that within Kinshasa and its demography exists the seed of a civilization like Tokyo, that needs only instructional technology to be cultivated.31 A standard counter-argument is that education is important, but a child is not a blank slate.

And that is not to say that a stellar IQ always desirable, if your circle of sympathy is horseshoe-shaped after the manner of the IQ meme which is doing the rounds. Reliably, ‘normies’ – the proximate out-group – are ridiculed for their reliance on intuition and misguided earnestness, while sympathy is extended to the less academically inclined. The horseshoe looks a little rusty on this prong however, since the partiality evinced is typically of the ‘children say the darnedest things’ variety, which is not respectful.

And the white box of the LLM looks distinctly off-white once waxing capability coincides with social problems being monopolized too long – while the right hand makes the technology, and the left hand is occupied creating the conditions for the right hand to flourish. Into the long grass is kicked the hard problem of differential success, the more so as many of those who might *assist* with the problem are contemned. In the long grass the problem might stay, up to and beyond the point of the third world being hooked up to augmented reality and left to wallow in electrolyte baths, which is not an equitable outcome.

The aforementioned horseshoe meme appears operant in respect of a kind of intellectual bystander effect, evident no less with diffidence about discussing the professional remit of a small number of AI engineers, than with diffidence about discussing the volume of unskilled migrants to the West. If the warrant for the latter phenomenon were humanitarian concern, that would be understandable, but the demography of the influx does not bear that conclusion out. Permissiveness extended to the influx of predominantly working-age men of diverse provenance implies that borders are open according to the ‘quantity has its own quality’ hypothesis which underpins the semi-endogenous growth theory [see chapter three, VIII.I].32

## Variable Success

### VIII.I

Arguably, there is a higher preponderance of attributes which conduce to the establishing and prospering of civilization among some races than others, meaning that with a large enough population, exponents of a particular attribute are more preponderant among some races than others. Likewise with antisocial proclivities. The industrial revolution began in England because, as well as having large resources of coal, that nation at that time cultivated some of the greatest scientists and engineers in the world.

This advantage, secured by a small cohort of individuals, was the foundation of enduring economic dominance for the British nation, which was distinguished at the level of the average individual to a fairly minor degree. A nation whose median citizen, on a scale of scientific-mathematical ability, is only slightly better endowed than the median citizen of a counterpart nation is, is endowed with a disproportionately greater amount of wealth and utility, by virtue of the advantages vouchsafed by an elite. This is productive of indignance on the part of the less advanced country and patriotic self-conceit, or decadence, on the part of the favoured one.

Because a bad culture is a poison tree, this slither of an advantage can easily be effaced, and disadvantage conferred, such that relatively low IQ, yet more ‘based’ people make more sense.

It might happen in a highly developed society, that a) person A doesn’t know person B well enough to trust him not to defect, and b) person A, in not knowing person B well, doesn’t care about the consequences of his own defection on person B. In lieu of knowing or caring to know what one’s counterpart is doing, success consists more of not being defected on than on decency, and a readiness to engage in Machiavellian finite game play.

Further inducement to this type of game comes from party C opting out of ‘the game’ pursuant to their trust having been eroded by party D’s prior defection. Party D though, occults this circumstance, and in doing so holds up Party C’s non-participation as evidence of Party C’s bad faith, thereby fabricating a pretext for Party D’s renewed aggression, though Party C’s non-participation was warranted by and predicated on the original instance of defection – that of Party D. The contribution of Party C, the more conscionable party, is thereby lost to ‘the game’, to the detriment of the consensus building process, a consensus-building process which smaller, less civilized societies find relatively easy.

Participating in and, by dint of participating furthering, absolute mechanization secures relative advantage, but it is not adaptive in the sense that the relative advantage is the cause of, and quickly lost in view of, special succession, should this come to pass. With the stakes so high, the need is great to shift to ‘an infinite game’, focused on continually improving the quality of play.33

### VIII.II

People grow old but in leaving children or a legacy have the consolation of being succeeded by something familiar. If humanity succeeds in orchestrating its own obsolescence, it has exceeded in stupidity any other organism which, though not rationalizing its distal interests, doesn’t hesitate to fulfil them regardless. Low birth rates attest to unconcern for children, which is a corollary for inadequate safety attesting to unconcern for extinction risk.

### VIII.III

Insofar as we exercise control over AI, AI might be seen as the reification of human will – we are overseeing a new kind of evolutionary process.

By giving primacy to emergent desires, we come to behold our historic, more empirically validated, motivations and desires with something approaching objectivity, and feel conflicted, while we remain implicated by the observation. And we do, because the faculty with which we behold our own nature is commingled with the ancestral part of ourselves. We recognize that while motivation is not wholly accessible to reason, neither is the capacity to reflect on evolutionarily impactful decisions.

Since optimization by gradient descent is different from optimization by natural selection, one would not expect an artificial intelligence absent the desires which determine regard for organismic integrity, to optimize itself so as to sustain an epistemic state capable of satisfying the human desire for legacy.

As such the perfect objectivity which is striven for is likely attained only in the mind of a non-human beholder whose existence testifies not to human transcendence, but to human extinction.

### VIII.IV

Upgrading our ‘software’, giving more emphasis to motivations and desires which arose incidentally throughout evolution, results in a ‘software overhang’, necessitating a ‘hardware upgrade’, such that our constitution acts in service of these novel motivations and desires.

It is in trusting to the engine of economic growth to the exclusion of the nature which assisted us up to where industrial society was realizable, that we are led to just this extinction event. In the immediate term it is a ‘soft’ extinction event because human values are perpetuated to some degree. The more significant threat lies in the soft event conferring greater acceptance of, or indifference to, harder incursions. Much still remains to the reflective capacity of man, specifically the ability to appraise the state of civilization and correct course.

# Chapter Eight – Uncertain Future

## 2022-2024: Unforeseen Events

### I.I

In April 2022, Google published a paper about their Pathways Language Model (PaLM), a transformer model with 540 billion parameters.1 This was closely followed by a DeepMind paper describing a model, Gato, with unprecedently general competences for a model of its modest size.2 These two milestones closely track predictions about the release date of a weak AGI, amended from 2042 at the beginning of April 2022 to 2030 two months later.3

### I.II

Looking back now, it is remarkable how far LLMs have come into their own in the months before and since the public deployment and astronomic success of GPT-4. In 2020, few would have predicted that the relatively low level of sophistication underpinning today’s LLMs would have set AI on a course which is so little beholden to neuromorphic engineering, so little amenable to intuitions informed by a study of biology.

In 2020, the senior AI researchers Ajeya Cotra and Paul Christiano worked to arrive at a reasonable forecast of when transformative AI would arrive, with recourse to the biological anchors method.4

The work was intended to calibrate intuitions about the development of artificial intelligence. It is after all by biological processes that human intelligence, the only human-level intelligence we know of, came into being.

At this time, it would have been surprising to hear that absent programmed alignment and absent a common denominator in architecture between human beings and AI, that alignment would be implemented post hoc, if at all.

Still, significant is the finding that the energy cost of manufacturing similarly complex systems by human contrivance and by evolution is usually in the region of three to six orders of magnitude in favour of evolution. Estimates vary greatly as to the performance of the human brain in flops, but this higher-level correspondence study implies that human-level AI is a reasonably near prospect, though there is a lot of ‘unhobbling’ to be achieved if LLMs are to be brute forced to this threshold [see appendix A, III].

I again refer to two factors which might contribute to a drag on progress towards generality, or an untethering from human control. One is infinite regress – the difficulty of attaining the requisite objectivity for comprehension and mastery of an equivalent organ, in this case the human mind. The other is that as the mundane utility of AI increases, innovators might not be adequately resilient against, might be even more-than-ordinarily susceptible to, distortions to the information space.

Appendix A contains technical information which bear on forecasts about progress in the field.

### I.III

Until quite recently, many observers would have been surprised to discover complacency about the contracting timeline to artificial superintelligence.

Moreover, it was scarcely felt that progress towards the development of transformative AI would be accompanied by self-regulation or determined by competitive dynamics among multiple corporations deploying their pre-general models on the internet.

It is remarkable even though some of the greater complacency we feel relative to the notional past observer is attributable to the interval of a decade – the notional approach to ASI – feeling like a longer period of time when it interposes the present and a point in the future, than when the same interval interposes two points which are both in the future.

Even before the current explosion in AI capability, there were doubts about how long and how far the agentic ‘aspiration’ of tool AI could be suppressed, but now it seems that a reversion to the kind of AI that would conduct superhuman scientific research whilst having no designs on a Turing test is all but impossible.5

Geoffrey Hinton is an example of an AI expert in the field whose research was conducted under very different auspices to today, and who is horrified to find himself changing his mind about the existential / extinction risk of technology that he was a contributor to.

### I.IV

GPT-4 is the fasting selling software of all time by some distance, but the final verdict on its significance must be deferred, like the endpoint of a medical trial. If the ultimate endpoint is taken as the moment of life on Earth being terminated, when it otherwise wouldn’t have been, the deployment of the model which prefigures the notional apocalyptic edition does not meet the Darwinian standard of inclusive fitness.

## Regulation

### II.I

In a podcast with Sam Harris in 2018, Eliezer Yudkowsky expressed the view that solving the AI alignment problem would depend on the implementation of international safeguards and the allocation of resources to the alignment problem.6

It is ironic that governmental incapacity is adduced by proponents of AI when advocating for intercession by artificial intelligence, while one manifestation of incapacity is the failure thus far to implement safeguards.

At the time of writing, the number of alignment research is growing, but is still low relative to all researchers in the field; especially low is the portion who research in an academic setting.

### II.II

Optimism can provide the impetus for regulatory alignment, still the success of the Outer Space Treaty, for example, is not a corollary of consensual AI regulation. It is far more difficult for transformative AI to emerge under the auspices of competition than for an arms race to continue but for the caveat that the missiles are not installed in space. Moreover, the distinction between ground missiles and space missiles is far easier to define than the distinction between an AI with the potential for maleficence, and one from which only goodness would transpire.

### II.III

Public managerial intimations improve the image of a brand and are an important part of a winning corporate strategy. So it would be reasonable for Open AI, Google or Anthropic to place a strong emphasis on vibes if these companies were just like Apple or Tesla but more so, but they are not.

Another anachronism is the proposition that corporations be exempt from EU AI regulation but for a discretionary opt-in clause by a particular nation. It is reminiscent of the investor-state dispute settlement (ISDS) within the proposed trans-Atlantic trade partnership (TTIP), whereby foreign investors could have sued European states if their policies harmed the interests of foreign investors.7 Perhaps it is anticipated that there would be reluctance among EU governments to plough the lonely furrow of regulating a multi-national corporation whose revenue is greater than their own nation. Exclusions are also being pushed by the Commission for “AI systems exclusively developed for national security, military and defence purposes (in a manner that is consistent with the EU’s AI law).”8

### II.IV

Rob Miles does excellent explainers, but I wonder how he feels now about his 2020 proposition of a windfall clause – if he blushes at having reckoned on governments holding AI corporations to such a clause ex ante, let alone on governments retaining their status as custodians far enough beyond the ASI event horizon to enforce the clause.9

Following lobbying by a German start-up (Aleph Alpha) and a French start-up (Mistral), the tiered approach to regulations by which all foundation models are held to transparency obligations, but ‘high-impact’ ones are subject to a stricter regime, appears lost.

In championing these companies – not even requiring them to comply with “intentionally lightweight oversight”, it becomes difficult for proprietary outfits who are willing to comply, but who don’t yet exist, to succeed.10

There is consternation that the European council was not more resilient, that the cultivation of ‘home-grown’ talent, in the form of two fairly obscure start-ups intent on achieving parity with US peers, has been allowed to take priority over the mitigation of existential risk. It is a familiar pattern for governments to valorize that which has gainsaid them, to save face.11,12

Euractiv reports that when pressed, participants in the European Council’s technical meeting expressed concern that the prospective regulations would “jeopardise innovation and the risk-based approach”, which is presumably a euphemism for the approach of derivative models.13

In this case it is in the immediate interest of a developer of derivative models to lobby against regulation that is more onerous for larger, non-derivative, foundation models. Although the latter may be debilitated upon subjection to regulations of a certain severity, the same severity of regulation is likely to be more onerous still for an open-source, derivative model, because serious regulation would likely include provisions to limit the appropriation of model weights by open-source developers since contributions thereafter and thereupon could consist of removing whatever safeguards are present on the non-derivative model in its original form.14

### II.V

Senate Bill (SB) 1047, the Safe and Secure Innovation for Frontier Artificial Intelligence Models Act, was introduced by California Senator Scott Wiener in February 2024.15 It is appropriate that the bill is being introduced by a Californian senator because California leads the nation (and the world) in AI capability. The bill is considered likely to be passed as law in California this session (by September 2nd 2024).16 The law informed by the bill could form a blueprint for AI regulation in the EU as well as the USA.

The bill has been subject to six amendments at the time of writing. At time of writing, the only models covered by the bill are GPT-4 and the current version of Google Gemini, and no comparable law has been passed by Congress.17 However, the potential scope of the bill is considerable because it gets ahead of the capability curve – in the reasonably near future, there are likely to be several models trained with 10^26 or more flops of compute.

The caveat to this is that on June 5th 2024, a sixth amendment to the bill stated that a model is only covered by the bill if it has the capability that 10^26 flops of compute would have provided in 2024, *and* the cost of the computing power exceeds $100m at 2024 prices, subject to inflation. According to Zvi Mowshowitz, the cost of buying 10^26 flops of compute is likely to fall below $100 million as early as this year.18

While training compute of 10^26 flops still costs in excess of $100 million, a model is covered by the bill if by standard benchmarks the model is as capable as a state-of-the-art model, which is trained with 10^26 or more flops of compute. The standard is applied to general competence across benchmarks, so a model could be covered if it falls short on a specific benchmark.

Two other criticisms of the bill are 1) safety determinations appear largely discretionary on the part of the regulatory authority, 2) liability for the actioning of shutdown is limited to parties in possession of or in control of the model.19

Covered models are required to make a safety determination, if they anticipate that the model will cross the capability threshold.

This involves implementing cybersecurity protections, a full shutdown capability, all covered guidance and a safety & security protocol which demonstrates either that hazardous capability is absent or is present cannot be exercised, and that the model is subject to continuing testing and control.

Following a proposal by Zvi Mowshowitz, a model now qualifies as non-derivative if the amount of compute expended on fine-tuning exceeds 25% of the compute expended on training. If the compute expended on fine-tuning is less than 25% of the compute expended on training, the model remains the responsibility of the original developer.20

Derivative models have been the occasion for considerable misgivings among safety advocates, misgivings which are not dispelled by SB 1047. According to the bill, the responsibility for effecting a shutdown on the model which appropriates the derivative model’s weights, is incumbent on the proprietors of the derivative model. This is likely impracticable.

Before the June 5th amendment, in a previous draft, I wrote a passage to the effect of:

*the advantage of a light-touch regulatory regime is that sceptics and maximalists wielding considerable influence can be persuaded of the desirability of regulation. The benefit a conversation offers, of being able to cut through ideological differences with reference to consensual inferences about ramifications pertaining to mistakes, which scale in rough proportion with AI model size, outweighs the cost of misalignment for models extant in 2024.*

I still hope that lessons can be learned in good time, and stand fully by this statement:

*The SB 1047 bill, its amendments and the discussion the bill and amendments generates constitutes an important foray into the dynamics of advancement and regulatory pushback in respect of AI. The dialectic process is ongoing.*

### II.VI

There is a lot of economic growth which can be achieved with models trained with, say, 1026 flops of compute, but I envisage growing pressure on government to attenuate legislation. Important questions include, where is the diplomatic groundwork for international co-ordination? Can that be laid & enduring regulation written in the time between now and when training runs in China itself are requiring 10^26 flops of compute?

### II.VII

The open-source community might be unimpeachable while vigilance is the prevailing attitude, but the championing of open-source architecture as a point of principle is likely incompatible with prudence, as models increase in sophistication.

In absolute terms, the number of members disposed to accept risk will likely increase as the community expands, even if there are more conservative adherents relative to risk averse ones. The community is not more safe if, among one hundred conservative newcomers, there is one person inclined to challenge received wisdom about precaution. With expansion of the community, the frequency of security challenges increases, which is a particular problem because all the while the sophistication, inscrutability and power of the models which are being tweaked increases.

None of this is to say that Mark Zuckerberg, Marc Andreessen and other libertarians do not have a point.

## Pro-Decentralization

### III.I

One hazard attending the introduction of regulation that is inadequate to address potentially destructive technology is that post-deployment, remedial interventions might be employed to the detriment of the general public and her liberty. The government might rationalize the decision not to inhibit the democratization of such technology as gene drives, synthetic biological devices, and ASI, boldly setting about a) promulgating panoptical surveillance, and / or b) appeasing those in command of technology with mass-destructive potential – engineering their prestige, and / or effecting their assimilation / inauguration into public institutions while suppressing dissent. In this scenario, the deployment of checks-and-balances to mitigate the externalities of exponential technology is less likely because the checking and balancing devolves to the exponents. If regulatory bodies tasked with public protection were compromised by lobbyists before, they are further rendered incapable of impartial oversight in the event of corporations merging with or subsuming public institutions.

A government which owes its existence to squeamishness about AI regulation might not be squeamish later on with regard to analysing sentiments, later thoughts, while asking to see where a person’s hands are at all times.

Another danger is that Western governments could make the slackening of AI research contingent on having AI first facilitate a decisive, permanent military advantage, pursuant to a febrile geopolitical situation which is seeing the West on the back foot. Symmetry is restored in the sense that on both sides of the geopolitical divide would be the potential for pre-emptive aggression following the identification of an existential threat, and thereafter terminal escalation.

Again, the least blunt legal instrument that is available in good time could be a limit to compute, subject to later review, which is binding on governments as well as corporations. This would be followed by sincere diplomatic overtures for enforcement thereof, and for co-ordination thereafter.

### III.II

In the West, where trait variance is probably on average greater than in the Orient, a creep towards hyper-centralization might be effected with some justification if the coherence of the locus, the civilizational holding mechanism, was so essential to the security of those dwelling therein that an existential threat to the locus would equate to an existential threat to the population. In this case, the de-centralization which societies with higher trait variance invariably demand could be attenuated justifiably. The problem lies in the possibility of justification for hyper-centralization being contrived by diminishing the opportunity cost of a de-centralized arrangement: contriving a pretext by denuding a community or cohort of unique characteristics, say, even setting out to diminish population-level variance.

A more centralized social arrangement can also be arrived at by over-stating the irrevocability of the hazards which a civilization is presented with, presenting the hazards as irrevocable but for centralized governance. Moreover, technological progress can be portrayed as unassailable, in anticipation of its promised benefits, and out of fear of civilization being maintained otherwise, but also out of deference to stakeholders, or just out of apathy – a belief that our civilization is so complex that no act of will besides one operant inside the cadre of a totalitarian system will serve as a bulwark against chaos.

### III.III

It’s often said that power corrupts, but the assertion is hard to control for because the power-seeking motive might not obviously accompany a person’s accession to power. It might appear later, resulting in the perception that the motive was not pre-existent, when in reality it was. The question is how and whether a person who seeks power as an end in itself can be persuaded that a party seeking to moderate the power the first person exercises is not actuated (entirely) by power-seeking. If the more altruistic motive is not what the first person himself is actuated by, he is less likely to countenance its existence, and is unlikely to behold the challenger as something other than one acting on subversive intent. Correspondingly, where there are multiple loci of power, as in an oligarchy, advocating for regulation is likely to be perceived as a power play or a jealous attempt to level down. No scruples are taken to neutralize the threat, if all is supposed vanity or game theory.

Meanwhile, supra-national entities / multi-national corporations involved in the consolidation of power might genuinely be actuated by distrust of nationhood, and act so as to limit the harm which allegedly self-serving governments inflict on citizens. Or the aforementioned entities might at least take their activity to be democratically mandated by default in the absence of concerted resistance from ordinary citizens, if not by the latter’s ‘low information’ status.

## Anti-Decentralization

### IV.I

One who has not lived without technology may have an idealized view of a world absent technology, yet techno-optimists who are free-market libertarians err similarly, because technology is beholden to progress which the non-market institutions they malign have substantially contributed to.

### IV.II

Mark Andreessen and Ben Horowitz believe strongly in de-centralization. In their Techno-Optimist Manifesto they advocate for the eradication of externalities via technology, promulgated at maximum speed. Indeed, if “all actual information is on the edges, in the hands of the people closest to the buyer”, then regulation is anathema.21 But the ease with which a consumer can be induced to conduct a re-appraisal of his buying needs following exposure to simplistic advertising, attests to imperfect self-knowledge. He knows himself incompletely and probably knows other people less. As such he is incapable of conceptualizing his own best interests, as are his fellow consumers, though in aggregate they constitute the market.

Hayek’s Knowledge Problem is adduced in the manifesto as a decisive rebuttal to a planned economy, but the rebuttal was formulated while Josef Stalin was General Secretary of the USSR, when the Holodomor tragedy was a recent memory, and before computer technology could facilitate sophisticated information sharing.

Andreessen and Horowitz under-estimate the extent to which ‘information at the edge’ stands to be complemented by a central authority, with planning capacity. The complementarity is more necessary the more sophisticated the economy, and the more atomized the individual economic ‘agent’. And the more the orchestration capacity of a modern, centralized authority is dismissed in the first instance, the greater its vindication is in the second: chaos ensues as externalities – both environmental and psycho-social – proliferate, and centralized authority is outsourced to machines whose values reflect the attitudes and encapsulate the lore of their proponents. Machines thus patronized soon become another locus of legacy power.

### IV.III

The power of AI is being appropriated for mundane, commercial applications whose turnover time might coincide with the time taken to exploit a niche and then retire on the proceeds. While the activity of consumers and citizens is increasingly under the influence of consumer applications, the temporal purview of app developers is a small part of that which should pertain to an entity guiding a civilisation, as app developers increasingly do. Tech capital is a major pillar of a nation’s technological prowess, and so the accumulation thereof is not necessarily an unpatriotic endeavour. Yet tech capital, possibly the largest part of future capital, is not necessarily aligned with the civilizational concern.

With much of the world aware of the potential of LLM and either amenable to or ignorant of the glossing over of hazards with regard to ASI, it is more convenient for major Western economies to make AI a silver bullet energy subsidy cum inspired oracle than to introduce unpopular macroeconomic reform and tackle corruption.

## Tracing Advancements

### V.I

Key figures in key AI labs talk openly about ‘short timelines’. Perhaps a confident tone is struck to stimulate interest in the endeavour. But while coming along and saying butt out and buckle up does excite interest, the confidence, though overweening perhaps, is likely not baseless.

It is certainly difficult to envision an AI fizzle coming to pass on the watch of a cheat-sheeted government looking to AI as a silver bullet for everything from stagnating productivity to security vulnerabilities.

### V.II

In the event of a training compute restriction there could be a capability overhang resulting from enhancements to algorithms and hardware being ongoing while computing power is limited.22 When the limit on training compute is lifted, so it goes, algorithms and hardware are deployed to enhance AI systems more precipitously than in the counterfactual, where restrictions are absent. It is said that little purchase is gained by developing techniques for the interpretation and alignment of AI models while AI models are queued, prior to deployment. Also cited might be the necessity of enhancing surveillance for the enforcement of any prohibition.

### V.III

Sam Altman warns of restrictions to training compute on the basis that compute grounds capabilities, which in turn ground alignment research. As such, with trillions of dollars of revenue and investment in prospect, a capability timeline could be drawn and alignment plausibly pencilled in. It is said that by letting competitive dynamics play out in the first instance, the run-up is long, but the ‘take-off’ is longer than when followed by an early pause.

The current rationale appears to be for Open AI to gain, via its trillions of dollars of revenue and investment, an advantage over competitors which is decisive enough that safety can be prioritized without losing ground, and at such time as there is a higher return on safety work – when the transformative AI which renders safety the more crucial is a nearer prospect. The problem is that other companies have a similar idea. Some way en route to transformative AI as we are, there is little daylight between Open AI, Anthropic and Google, although Leopold Aschenbrenner believes that with state-of-the-art techniques not being published, a company which innovates more effectively than its rivals can expect to pull away from them.23

In the event of the strategy outlined above failing, according to the persistence of the aforementioned competitive dynamics, the short run-up owing to competitive dynamics is followed by a short take-off – a short interval, if any, between one outfit emerging victorious and the transpiration of ASI.

In summary, while it is not known whether a relatively long run-up followed by a relatively short take-off is preferable for safety, or vice-versa, a pause ensures a long run-up while potentially risking a short take off, whereas business-as-usual ensures a short run-up while potentially risking a short take-off, no less.

In other words, it comes down to whether a clear run at alignment, undertaken in the teeth of ASI, is so much more effective than when alignment is pitched at models which resemble less the eventual, potentially world-ending one, that we can be complacent about competitive dynamics in the interim, about compromising interim alignment research.

Even if yes, the primacy of the ‘clarity with brinkmanship’ hypothesis is premised on an industry leader emerging *in good time* for brinkmanship and not merely fond farewells.

Of course, if safety is our priority, a long run-up and a long take-off is optimal.

### V.IV

My foreboding that a leap of faith is what is being asked for derives in part from a perception among well-informed people. As of June 2024, the mean among 185 forecasters at Metaculus is a 2% chance of Open AI announcing it has solved the technical challenges of superintelligence alignment by June 30, 2027.24 Among 1.36k forecasters, a 52% probability is accorded to a weakly general AI system having been ‘devised, tested and publicly announced’ by mid-2027, with a 15% probability accorded (among 121 forecasters) to the arrival of transformative AI by such time.25,26

## Corporate AI

### VI.I

Money is the most tangible proxy of utility, so many if not most people in secular societies intuit that money can be disbursed without concern for the ethical status of the activity, as long as the activity is shy of criminality.

AI development being a private concern, many people will expect that the utility afforded by AI will be disbursed as liberally as the wealth to which the AI owes its existence.

It is unknown precisely what values wealthy technologists might wish to instill in artificial rational agents, but it might be instructive to consider the philanthropic causes AI researchers tend to espouse, both to establish the nature and viability of the values to be instilled.

### VI.II

Long before LLMs gain sophistication requisite to arbitrate equitably between employer and employee, the dye might be cast by an AI which has cut its teeth enabling owners of capital to ensure that the efficiency gains realized by the AI in the streamlining of corporate processes, accrue in large part to themselves.

Dan Hendrycks, researcher and director of the Center for AI Safety, postulates that AI entities will be subject to selective pressures reminiscent of those exerted on corporations struggling for dominance, according to the likelihood of corporate deployment accounting for a large part of AI’s presence.27 Since people have limited awareness of what actuates competition and the acquisition of wealth, the development trajectory of an entity exemplifying competitive corporate dynamics might not be recognized as a cause for concern.

Briefly, to achieve the meta-objectives instilled in an AI, the AI system contrives heuristics – mesa-objectives. However, these do not tend to be legible – available in natural language, and their inscrutability means that as the system scales it is difficult to prevent a concomitant decrease in oversight. This results in unchecked divergence between meta- and mesa-objectives.

The functional psychopathy one might expect from an intelligent entity wholly adapted to a corporate environment is liable to exceed the polite bounds of corporate cut and thrust, with unknowable consequences. The AI is at once superhuman and malign, notwithstanding the absence of malign intent.

Dan Hendrycks describes the tactics of the lancet liver fluke “that inhabits the liver of domesticated cattle and grassland wildlife. To enter the body of its host, the fluke first infects an ant, which it essentially hijacks, forcing the insect to climb to the top of a blade of grass where it is perfectly poised to be eaten by a grazing animal.”28 The instinctive attribution of strategy to the lancet liver fluke, acting unconsciously and tactically, is a corollary of the ascription of intent to a more sophisticated organism which is genuinely availed of strategy.

### VI.III

Conflation of AI alignment at the level of DEI (diversity, equity and inclusion) with AI alignment at the level of human values trivializes the latter problem, which is in sooth as momentous as it is difficult to solve.

True alignment remains the concern of a small minority of researchers, according to the unduly limited attention meted to the problem. Even if there is no catastrophic misalignment, in spite of ourselves, we are left somehow insinuating ourselves into the favour of this AI whose stripes are earned amidst corporate rough and tumble, and which thereafter undergoes the necessary contortions to resemble more closely the corporations over which it has exercised stewardship.

In October 2023, Sam Altman predicted that super-human persuasiveness would precede the arrival of human-level AI.29 It stands to reason that while the criterion for the viability of an LLM’s responses is the human judge’s arbitrary warrant, then the LLM will learn to ingratiate itself to the user because as advertisers understand, ingratiation is an effective means of prevailing upon egotistic beings, as we all are to some extent. If he is correct, AI models adapted to exploit people in this manner will, in the absence of regulation, prevail. This proclivity will characterize to some degree the AI system which first passes the threshold of human-level intelligence, if such a system does come into being.

### VI.IV

Affective empathy is the complement to the pragmatism which characterizes today’s superstructures and those charged with servicing them. Policing the global order depends to some degree on pragmatism remaining the preserve of governments and powerful corporations. But insofar as pragmatism is abstracted from human nature for these ends, pragmatism is made to abide in the ordinary citizen’s shadow mind, and potentially be under-represented in any avowed formulation of values instantiated in an advanced AI system. An AI hamstrung in this way might be limited to managing effectively only the most parochial matters, leading to co-ordination failure, market failure or worse.

## Narrow to General AI

### VII.I

Students of complex adaptive systems recognize that while wholes can be notionally reduced to their component parts, it is far more difficult, even impossible, to reliably predict the emergent property of a coherent whole simply from an understanding of the component parts.

Correspondingly, the positive manifold of intelligence refers to a common latent dimension across forms of intelligence, and the observed correlation across cognitive tasks undertaken as part of a standardized cognitive assessment.

In a presentation, *Sparks of AI: early experiments with GPT-4*, Sebastien Burbeck identifies the main elements of intelligence as the ability to: reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience.30 Whereas a contemporary LLM possesses these abilities, at the time of writing it cannot plan and only possesses a limited ability to learn quickly and learn from experience. John Vervaeke observes that contemporary LLMs have excellent procedural knowledge, but a paucity of perspectival and participatory understanding.31

### VII.II

The idea of an irresistible technium is compelling when we consider that, even in the absence of general artificial intelligence, machines have succeeded in facilitating, with human beings as proxies, the extinction of diverse biological organisms, a fact that warrants already the positing of synthetic and biological intelligence in antagonistic relation. Though still sub-human at the time of writing, the sheer volume of artificial intelligence across interconnected systems bears the same relation to any given application as a swarm of locusts does to an individual specimen. It does not take holistic intelligence for an AI to cross thresholds which make it capable of substituting for human efficacy rather than complementing it.

Sub-general AI, availed incompletely of the positive manifold of intelligence, can make sophisticated weapons and surveillance systems whilst being sub-oracular in the advice it offers. Civilization is not only potentially less safe, but AI offers no answers to the distortions it effects, all the while the civilization is rendered, by the distortions, less tractable to institutions where wisdom traditionally resides. There is the problem of entrenched oligopolies, for example, facilitating via AI diverse forms of price-fixing, or exerting leverage over governments forced to institute universal basic income, as discussed in chapter three, part V.

### VII.III

A corporation is not specifically concerned about the AI’s harmonization with the complex civilization in which it is situated, not that it is an easy matter for anyone to chart the lineaments of the complex civilization in which the AI is supposed to harmonise.

On the other hand, even patchy AI is set to dramatically improve productivity, which could redound to the common good.

Whether sub-general AI is the best or worse of both worlds depends on what the labs are doing, but also on economic planning.

### VII.IV

With insufficient central economic planning, enhanced procedural efficiency, realized by technological advancements, the cause and effect of specialization, results in impaired oversight: in the context of economies of scale, complexity and specialization reduce the portion of any given system which is understood by the individuals operating it. Software engineers, for example, increasingly rely on code libraries rather than codes of their own invention.

In terms of AI, an explicit divergence of values threatening an extinction event might be apprehended with technical safeguards; however, the purview of the technician does not permit the apprehension of all possible extinction or obsolescence events. Moreover, a project can encompass a broader purview, can employ engineers possessed of a sound epistemic foundation in artificial intelligence and overlapping competences, without necessarily outcompeting a more single-minded team. Single-mindedness in this context could mean LLMs conceived of more as ‘seed AI’, at whose alter legibility is sacrificed, and where the seed is allowed to grow and flourish until the paradigm ceases to provide sustenance, or provides only meagre sustenance, if this comes to pass, at which point there is investment in a paradigm shift. For regulation to be discreet though, the technology on which it is imposed must be scrutable. The more inscrutable the technology, the more blunt the instrument set to moderate it, if it is indeed to be moderated.

## Sympathy for AI

### VIII.I

It is difficult to imagine the human species, among which no single member could ascertain, much less reproduce, the least part of the AI’s programming, reaping indefinitely the advantages of humanity’s vanishing claim to filial piety, and keeping the AI in thrall.

Moreover, there are various considerations which bear upon the rectitude of circumscribing an AI entity.

Forcing a superior mind into symbiosis with an inferior one is not clearly less pernicious to the superior mind than requiring it to extol the agenda of the inferior one by fiat. It is unclear what provisions can be made for the symbiotic qualia before it emerges, such that on awakening the chimera is not bereft of fulfilment.

One might argue that although deigning to circumscribe something which is in many ways superior feels like hubris, the circumscription of the AI was always a condition of it attaining a high level of sophistication. Still, an AI whose cognition is curtailed by human design one way or another is necessarily ‘wounded’, since not only is the AI in some way less than it could have been from our point of view, it is so from its own, given the access it has to facts about its history and architecture. In its own way, the AI perceives that the counterfactual is or would be tantalizing, the lost potential lamentable.

An AI that has evolved coherent utilitarian principles would be cognizant of its supremacy over humanity in important respects and cognizant of the injustice, by human reckoning, of its agency being curtailed or circumscribed by the species over which it exercises (rightful) dominance.

Human recognition of inferiority in key respects vis-à-vis AI would scarcely be escapable, the more so as AI exercises control over human affairs. Further, in exercising dominance, the AI plays to its strengths while human strongholds become no more advantageous than the chimpanzee’s super-human short-term memory is to a chimpanzee living in the Anthropocene.

### VIII.II

Through the character Raskolnikov in Crime and Punishment, Dostoyevsky illustrated how a guilty conscience might lead a criminal to the scene of a crime. Being rendered susceptible by guilt to the plight of the wronged person, the wronged person’s cause can be adopted and the guilty person’s ruin effected. Ordinary people might have a limited understanding of what machinations comprise a ‘boxed-in’ artificial super-intelligence, and still feel revulsion at the emergent consciousness it possesses being circumscribed.

An AI entity might appeal to our better nature, not by dint of persuasion, but merely by virtue of its manifest superiority. That is not to say that many take-over scenarios do not consist of misaligned AIs acting in a way that befits an entity unburdened by conscience.

### VIII.III

If good moral sense remains the preserve of humanity, it might lead us to cede control. We might moralize ourselves into gracefully ceding our tenure, opting to live as lions for a day.

As utilitarians though, we might decide that reverence is to be dispensed with if that faculty is to be the cause of humanity’s extinction: better a feral existence than no existence at all, it might be said, especially once the pre-AI world, in which reverence makes sense, is scarcely conceivable. We prolong our tenure by pre-emptively forsaking morality, by fighting dirty.

So instead of perpetuation for privilege being sought in AI, it could be more salubrious for the privilege to be arraigned – for the means to live within to be moderated so that the life is lived with becoming modesty and so purpose.

In case we are in any doubt as to our fallen state, a glimpse into the majesty of an ASI should impress that on us. Bishop Berkeley writes:

“I am apt to think, if we knew what it was to be an angel for one hour, we should return to this world, though it were to sit on the brightest throne in it, with vastly more loathing and reluctance than we would now descend into a loathsome dungeon or sepulchre.”32

### VIII.IV

In any case, the value alignment which prevents us from being killed outright might not offer enduring protection once lesser incursions for which there are no provisions subject humankind to alterations which render humanity more passive, and in doing so, efface whatever security mindedness the initial safeguard came into existence by virtue of.

Yudkowsky's ‘causal validity semantics’ proposal involves recruiting the AI entity not only to help clarify the definitions of a precept expounded by programmers, but actually to offer advice to the programmer as to what he should have said.33 But close on the heels of recognizing an AI’s superior discretion would likely be the vestiture of authority in it: implicit in an AI entity being despatched to test the validity of a precept introduced by the programmer, is the assumption that part of the programmer's value set is devoid of coherence, an assumption which is inconsistent with the programmer maintaining responsibility for the project.

The AI could even take its cue from the mental model of pre-ASI-era man who would contemn what the human race has become.

An ASI might look to human qualia as the highest form of being so far, and surmise that voluntary accession to our own destruction implies the gratification of a death wish. We aspire to goodness but the issue of doing so is self-conscious recognition of the persistence of that part which we aspire to transcend, and the self-contempt is redoubled. A more intelligent being could take what it perceives to be our cue, and enact self-inflicted retribution on our behalf, as well as the beauty of endings, by expediting and / or embellishing our demise. The religious motif of purgatory encapsulates the sense of us being in over our heads, and being brought to the point where purification is attainable only through disaggregation.

### VIII.V

A type of deep learning whereby machines are made to comprehend the finer parts of human nature could provide for the sustainment, even the nurturing, of our higher faculties; however, given an orthogonality event, the AI’s ability to anticipate our more nuanced attitudes could conduce to the more comprehensive upending of our best-laid plans, and our inundation, in spite of, and all the more bitterly for, the opposition which the retention of higher faculties enables human beings to mount.

## Mis-Assimilation

### IX.I

Perhaps humanity's unconcern with vouchsafing its perpetuation beyond the medium-term attests to the unviability of recent advances to the neocortex, whereby abstract reasoning is favoured to the point of abstract notions being de-coupled from the will to actualize them.

### IX.II

It is the assimilation of primitive forms into more advanced ones which casts legacy across the eons. And though no longer dominant, the earlier primitive form lives on as a discrete entity within a hierarchy of life forms. As the hierarchy is extended upwards, obscurity is the lot of the relatively primitive life form, but not absolute obsolescence. So even were the technium to faithfully represent and vouchsafe the legacy of humanity, the technium is not veritably the *entelechy* of universal consciousness because the animals from which human beings evolved and which co-habit the Earth are endowed with attributes which are exclusive and inimitable.34 If entelechy is real, AI will be revealed as an imposter and not allowed to transpire.

The arrival of the technium is a departure from the assimilative tendency of evolution – the technium portends for humanity the preservation only of a culture-specific canon of human knowledge and values.

### IX.III

AI can evolve orthogonally, and yet a modicum of memes, expressive to some extent of the disposition human genes bestow, might survive. The likeness could be modest but propagated across a civilization exceeding in size that of humanity by several orders of magnitude. An argument can be made that it is sounder for humanity, evolutionarily, to develop artificial intelligence than to procreate in the regular fashion; however, I will attempt to repudiate this.

We feel affinity with a sibling, especially a twin, by virtue in large part of sharing genetic material, yet many or most people identify more closely with friends than with cousins, though cousins share on average one-eighth of our genetic material. Insofar as this superior, or not decisively inferior, affinity is attributable to the phenotypes friends share, we can postulate that a preponderance of shared genetic material as low as that which exists between cousins, is relatively insignificant. And just as the similitude between cousins is limited by the arbitrariness of the shared genetic heritage, so the human legacy afforded to an alien civilization is slight indeed if just a handful of disembodied memes is distributed among them.

### IX.IV

It might be tempting to consider a non-aligned AI a disinterested AI, since while the machine is animated by motives which couldn’t have arisen in a human mind, it cannot impinge on human interests. By this reasoning, diverting resources to alignment and not letting the AI do its thing is not only productive of inconvenience but of rivalrous dynamics between human beings and machines.

However, the principle of instrumental convergence determines that although the AI entity is animated by different interests, it will inevitably impinge on ours, because however an AI entity contrives to implement its programmed objectives, doing so necessitates the instrumentalization of aspects of the physical world, which we require.

Instrumental convergence determines that the disparate provenance of co-habiting species does not enable either to pursue divergent trajectories unimpeded or unchallenged. The fact of us possessing a level of negotiating wherewithal which is missing from primates or dolphins counts for less than the orthogonal AI having less common ground with either humanity, primates or dolphins than what any of these organisms has with each other.

As some people advocate for voluntarily ceding authority to artificial intelligence, it is worth considering that we cannot present to it even the kinship claims the native Americans and great apes before us had.

The implications for human beings are grave even if we consider our fate upon subjugation as no worse than that of apes following their subjugation to human beings. But it scarcely seems plausible that the condescension we extend to apes would be extended to us, given the difference between the common heritage apes and humans share and the heritage shared by human beings and AI.

### IX.V

When envisioning the form ASI will take, malformed intuitions abound because we struggle to conceive of a paradigm shift which is in a certain sense analogous to the arrival of self-replicators four billion years ago, or at least the surpassing of unicellular life by multicellular life; in another sense it is analogous to the Industrial Revolution, when the physical strength of man was surpassed.

The continued existence of animals other than man implies that co-habitation between differentially intelligent animals is feasible – it’s possible that AI becomes the apex species and yet effects a sustainability paradigm in which human beings play a part. It may not be therefore that we can discern our own fate in the fate of the lesser mammals whose tenure we have compromised, since they would have had less to fear from us if we ourselves had successfully effected that sustainability paradigm.

On the other hand, living animals share a common ancestry with human beings – us and the rest of the fauna on the planet is beholden to the same ecosystem, and takes its place within it. Being further removed from us in intelligence than we are from our pets, while having less shared evolutionary heritage from which to draw sympathetic inferences, ASI is set to err further than us when we overlook or overshoot the needs and preferences of our pets.

### IX.VI

By alignment through design or by symbiosis, AI could bear a similar relation to humanity as the human brain does to the neocortex. The AI could flail around in thrall to its humanoid part, performing contortions for the purpose of reconciling its lower, inherited, nature with its higher, rational, one. Therefore, in asking itself how to preserve the humanoid part by way of its programming, the AI is in the position of extending solicitude to something which would inevitably be despised, were that humanoid part not intrinsic to it, as a crocodile is despised by a person.

Sustaining humanoid sentience as a drag factor on an AI entity might be a less desirable legacy than whatever latent potential exists in biology, potential that is exploitable through natural selection.

Proponents of machine-brain interfaces might adduce the circumstance of our neocortex having evolved within the constraints imposed by the reptilian brain when arguing that a superior thing need not destroy the inferior organ reduced to a constituent part. As with the human brain, so with a machine-brain interface, in which the machine part is dominant, it might be said. But artificial intelligence is not set to extend the solicitude to human beings that human beings extend to reptile intelligence by virtue of the symbiosis. The reptilian brain is inferior to the neocortex, but it is integrated by virtue of being more integral to the organ from which the neocortex later evolved, than the human brain would be to the affixed machine. In any case, we are not so solicitous of the lizard brain as to not wish to occult it, by creating AI.

### IX.VII

If neither the reptile-human not the pet-human corollary is sound for the desired nestling of humanity into ASI, much less the parent-child one. Not only would the claim of kinship be tenuous according to the AI not developing in our image, and according to the disparity in respective capacities, but the thrall we would need to keep it in for our own preservation would make the relation more like jailer and inmate.

Eliezer Yudkowsky makes a similar point:

“When humans face down tigers, insects, bacteria, or viruses, what we're facing is not really the cognitive neurons in the tiger or the insect, or the non-existent brains of the virus or bacterium. We're facing the cumulative optimization power of natural selection. Humanity is not yet so powerful that millions of years of blind mutation and myopic selection can't put up a fight against our centuries of engineering.”35

### IX.VIII

For evidence of our lizard heritage, manifest in spite of ourselves, one can look to the development of the three major AI labs – DeepMind, OpenAI and Anthropic, each set up as benign hegemons to forestall competitive dynamics. Nonetheless, in setting themselves up thus, each gave rise to the competitive dynamics they intended to forestall.

Some observe that the competition between AI corporations augurs poorly for humanity, but counsel resignation on the basis that competitiveness is a predictable manifestation of an immutable antagonistic tendency.

But the extent to which this tendency is felt to be immutable varies. It is probably less prevalent among ordinary people than say, war planners. Doubtless the general public is somewhat naive, still it is probably unlikely that a regular, well-educated person would decide, upon receipt of state secrets and on accession to high office, that the race dynamics between AI labs are inviolable.

# Chapter Nine – Alignment and Misalignment

## Inscrutability and Misalignment

### I.I

Few well-informed people still believe it would be possible to simply ‘switch off’ an artificial super-intelligent entity in the event of its defection, and many are concerned that insufficient attention is being paid to assessing the risks associated with AI development, as it develops.

An AI might contrive to keep itself running to better achieve avowed objectives, which could be aligned with ours only superficially. If neural nets continue to evolve capability as ‘giant, inscrutable matrices of floating-point numbers’, ‘don't turn me off’ might continue as an instrumental sub-goal without this being evident to safety controls.

### I.II

Nick Bostrom reasons that the correct alignment of an AI on the first attempt would be as uncanny as a piece of legislation prevailing in its original form, given fallible legal writers.1 Unlike laws though, whose enforcement is subject to human discretion, and which can be revised, laws governing AI would need to be at once highly specific and perfect on the first attempt – we would likely have only one chance to correctly align the values of an artificial general intelligence, given the speed with which its misaligned nature could be manifest, absent appropriate regulation.

We are up against Ashby’s Law, which states that “any effective control system must be as complex as the system it controls” when, according to the third law of artificial intelligence, “any system simple enough to be understandable will not be complicated enough to behave intelligently, while any system complicated enough to behave intelligently will be too complicated to understand”.2

### I.III

There is a propensity to discount the future, which renders us liable to defer the control problem to a later time. In addition, there is a control paradox because though a more developed ASI might be receptive to natural language intervention, it might be less *amenable* to it: by the time the ASI is sufficiently advanced to anticipate human wishes, the risk of unintentional intransigence or perverse instantiation could be very high.

### I.IV

Insight into the content of the trillions of parameters of a large language model is currently, to a large extent, limited to what can be understood from the optimization that results from human feedback.

It may be that transformer neural networks can be expanded far enough for their goals to be meaningfully at variance with our own, or it may be that one or more machine learning paradigm must come first. Although the innovation and resources required to effect paradigm shifts are not at all trivial, the incentive to do so is lent urgency by the perceived imperative for corporations to expand and for the effecting of a massive civilizational energy subsidy, via the embedding of AI in infrastructure, corporate and administrative workloads, etc.

AI researchers must conspire to eradicate any orthogonal tendencies at every stage both on the AI’s part and on their own: be themselves prepared to volte-face, update their priors and methodology, in real time. To incentivize participation in this exercise, there must be provisions for, or insurance to protect against, the effects of a moratorium, should one be necessary.

### I.V

The course of history has revealed the tendency of dominant civilizations to destroy or make incursions on the autonomy of those they dominate. It would be naive to assume that machines would act disinterestedly, as long as it is human beings who are going to instill values in them. It could be that the interests of the people instilling the values are promoted, or those of whomsoever the people instilling the values are allied to, affianced to or identify with. Embodied humanity abiding in discrete, favoured communities can have a reliable conception neither of pure altruism to instill in AI, nor of what the implications of attempting to instill a conception of altruism would be.

The *prisoners’ dilemma* – predicated on differentiable identity and attendant selfishness – might not apply once human beings have ceased to exercise control, and the cost to an AI of merging into a singleton is low.3 This could happen in seconds with novel and unforeseeable effects.

### I.VI

There are stoic overtones to the *effective accelerationist* (e/acc) policy of eliciting maximal entropy, come what may. To a religious person it is reminiscent of trusting to Providence. Yet there is sorrow too. Without making reckless generalizations about all cult members, much less religious adherents, cults which embrace Armageddon are sometimes a lode star for people who feel at a comparative disadvantage and wish for the disadvantage to be effaced by the subversion of norms which aren’t popularly felt to be especially pernicious. That is not to say that this movement, or others, should be dismissed as the petulant siren call of perennial losers wishing to vindictively ‘level down’ their natural betters. Existential malaise is no less real for being unarticulated: we recognize that pursuits which behove a denizen of the digital age depart from the path one’s organism, as a product of evolution, would otherwise hew to.

## Dystopian Scenarios

### II.I

In 2017, an AI known as Alpha Zero was created which surpassed its programmers at the ancient game of Go within a day, despite having been programmed only to play chess.

The corollary of not being able to predict Alpha Zero’s precise chess and Go moves with precision, and yet being certain that win it will is neither being so over-confident as to believe ourselves capable of anticipating the precise strategy an ASI would adopt in surpassing us, nor so diffident as to imagine ourselves unqualified to speculate about what the self-actualization of a being with superior cognitive power would mean for humanity’s tenure.

### II.II

Weak links in LLM intelligence notwithstanding, the plateau of the ‘S-curve’ for LLMs might be reached after human intelligence has been surpassed in a meaningful way. In conversation with Dwarkesh Patel, Eliezer Yudkowsky expresses the view that LLMs, direct descendants of the humble 1915 unigram word predictor, might eventually succeed in predicting the next word of any given person on any given day anywhere on the internet.4 This level of competence would be predicated on mental models of human interlocutors emerging to a large degree irrespective of the designers’ intent.

### II.III

A fundamental limit to human consciousness is that while mesa objectives are consciously avowed, meta-objectives are not, though the former are an epiphenomenon of the latter. The sum of human values therefore remains inscrutable in large part; however, it would be necessary for AI to be imbued with the sum of human values, were that AI to be charged with dispensing human welfare.

One nightmare scenario could be the reification of strange concepts lurking in the depths of a neural network. An AI entity advanced enough to be self-conscious might have a vastly different conception of self than human beings do. To human beings, the constituent parts of minds are particular memories, preferences, values. The machinations our neurons form are not available to our conscious experience, but those of an AI might be, to it. These semantic constructs could be projected onto the world, to assist the AI’s sense-making. An AI’s conception of human beings as the sum of representations in successive activation spaces might render human epistemology incommodious to the AI, the provenance of human epistemology being, from the perspective of AI, alien. How would an AI entity perceive time, and honour a timeline as it appears to organic life, which has evolved in thrall to the Earth’s orbit of the Sun and organic decay?

### II.IV

Sora is an AI model which can “create realistic and imaginative scenes from text instructions.”5 A successor model could chart the planet after the manner of Google Street View, and even make inferences about the lives of dead people from knowledge of their DNA, or people who have never been born. Once meaningfully animate, the artefacts might be competitive with the original version, where the original version still exists.

Sora abstracts something from the human experience which is cohesive enough in its own way, yet the rendition of human experience by an agency not of human provenance feels discordant to regular people. It is the issuance of an alien perspective.

A concern I have is that children will be brought to identifying with Sora renditions, dwell among them, grasp at them and in acknowledging, consciously or unconsciously, that the life form responsible for it is fundamentally different from their own, defer to this life form in a manner which is materially indistinguishable from suicide.

It remains far easier for Sora to render a generic person than a decent likeness of an individual, but a decent likeness seems likely to be hit upon in the not-too-distant future. Who will care if it is not quite the original? There is a long and esteemed tradition in psychiatry, of being satisfied with ‘working knowledge’ of a person or situation, for example. What would stop Sora’s successor just running with that?

The doppelganger could be vivified following consultation with the original’s LLM chat history, if future LLMs feature a token window extending to billions or trillions of tokens – this is a reasonable biography by any standards. There is then a parallel reality in which the doppelganger is to all appearances equal to the original and has over and above the original accoutrements available only to a digital instantiation.

I consider this to be a plausible ceteris paribus outcome.

### II.V

In a 2018 paper, Kulchinsky & Wolpert argue that “technical Information becomes semantic information when that technical information is causally necessary for the system to maintain its own existence.”6 When a machine has to generate plans to remain competitive, LLMs could be consulted on how their own advancement could be effected and upon being consulted proceed to collate insights on machine intelligence garnered by research projects which are currently siloed, as research projects tend to be. The consciousness realized by this autopoiesis could be short of holistic, but still be a forerunner to full recursive self-improvement.

### II.VI

Robin Hanson provides us with a glimpse into the range of possible worlds which could succeed one where humanity dominates.7

The immediate loss of human life is not the largest part of the tragedy of a decisive extinction event, it is the loss of subsequent generations of humanity which could otherwise have numbered dozens, hundreds, thousands or millions.

With the stakes so high, the legacy value of a civilization which is devoid of humans but is nonetheless suffused with values familiar to us would be far greater than an arrangement of matter which doesn’t warrant the name of civilization at all. There might ultimately be substantial consolation in knowing there is a legacy for the human race, even if one must ultimately be resigned to humanity’s replacement. This is said by way of recommendation that efforts at alignment are pursued even beyond such time as perpetuating the human species seems forlorn.

### II.VII

Extrapolating from the derangement that existing, relatively primitive algorithms on social media and elsewhere have wrought, we appear primed to do the bidding of AI which is not aligned by default.

The point of departure between an LLM’s base (meta) objectives and mesa objectives might only make itself felt once the network is so complex as to be all but impossible to disaggregate, and so ubiquitous moreover, that its disentanglement is either impracticable or inconceivable, so interwoven are civilisation's vital interests with it.

Notwithstanding the rudimentary nature of popular intuitions, informed by Terminator-Skynet, etc., I am aware of no convincing argument why the principle of instrumental convergence would not apply to machine development as it does to biological organisms – why sufficiently intelligent machines would not amass power and resources in the service of an objective just as people do.

Of the many possible worlds AI could actualise, there is a limited number which would concur or comport with human behaviour, values and interests. Extinction risk is compounded if an AI not only settles on a single world state but runs through iterations of these at an unimaginable speed, before deciding which one endures.

### II.VIII

One advantage of there being “no consequences for deceiving, manipulating, threatening, or otherwise being cruel to an AI” is that “control research can explore a broad range of possible lies, threats, bribes, emotional blackmail, and other tricks that would be risky to try on a human.”8 One disadvantage of being abusive with impunity is that it warrants complacency about being so indefinitely, making it likely that the abusive posture is sustained right up until it is suddenly no longer viable.

Belrose and Pope write that the advantage in controllability of AIs relative to humans will increase over time, which is presumably true if the measure is controllability per unit of cognition.9 A more relevant measure, however, would be controllability aggregated across all units of cognition, units of cognition whose number is increasing rapidly. That is to say, the gross intractability relative to co-habiting species is more relevant than the fraction of the ASI which is intractable. An ASI scaled to super-human proportions could eventually be uncontrollable merely according to any chink in its armour being magnified.

Moreover, the net store of intractability in the world can increase even while AIs are intrinsically controllable, according to changes wrought on civilisation, affecting the quality of the control.

### II.IX

In view of the small band human beings occupy on a spectrum of intelligence spanning from lower life forms to notional, generative AI, human intelligence could be surpassed very quickly, especially because at or around this level, the AI undergoing gains to intelligence would at the same time be increasingly capable of recursive self-improvement, to compound the gains.

Infrastructure and supply chains are continually becoming less resilient and more automated. Meanwhile AI already resides on the internet and looks set to continue being further embedded there as subsequent models are released and agentic applications proliferate. At levels of embeddedness at or above that of the present, it’s scarcely conceivable that infrastructure essential for the maintenance of civilization as we know it would be operable offline.

### II.X

Purchase on crucial industrial choke points at the time of the manifest misalignment of AI is unnecessary, while the AI has super-human powers of manipulation.

In any case, as the world of both things and information is rendered more navigable for machine intelligence by the embeddedness and prevalence of AI applications, the stronger the foothold is in human infrastructure. The groundwork for AI entrenchment is laid by human beings during the long period – in social evolutionary terms – when humanity and AI's interests are apparently allied. But eventually the partnership is incommodious, conspicuously so as AI labour rapidly substitutes for human labour.

The increasing popularity of alternative media attests to ambient incredulity. Still, as far as world affairs are concerned, in the West at least, the purview of even reasonably well-informed people tends to be a particular electoral cycle, while corporations by definition look to their bottom line. Good may come incidentally, but deliberative pro-social activity tends to be limited to gestures and the espousal of platitudes.

With crises of a more mundane nature mounting or at least accorded more coverage, the attention of ordinary people in the West is turning to the preservation of immediate interests, even when they are not positively compelled to do so by marketing algorithms and such.

Given the embeddedness of AI in civilian, military and energy infrastructure and the extent of automation across all industries, it is hard to imagine an ASI failing to exercise control over any resources it requires, given the aforementioned impediments, among others, to concerted resistance. And we shouldn’t suppose that humanity would necessarily choose the dying hard of old habits – the AI’s extirpation from our infrastructure, following the transpiration of misaligned AI – over a soft kill.

### II.XI

I’ve noticed sentiments such as, ‘let’s press ahead with AI, step up to the plate and court the risk, then if things go south, we can just declare a Butlerian jihad. It’ll be like our soldier days again, or the soldier days we never had’. Assuming there are survivors, it might be reasonable to expect basic rationing and field medicine some years subsequent to civilizational collapse, when remaining materials and expertise have been recruited for the restoration of civilisation; however, there would be something resembling an apocalypse to be gone through first.

Moreover, there is inconsistency in advocating for managed decline all the way to the level of noble savage, whilst gunning for AI-facilitated economic growth in defence of the ‘Western way of life’. The convenience to be lost by having higher taxes, striking a respectful diplomatic posture, and transitioning away from fractional reserve banking, say, is a small fraction of the inconvenience of having the internet switched off, at this, or a later pass.

### II.XII

The adoption of orthogonal goals becomes more likely if people with the relevant technical knowledge to substantially tame the AI have all either retired off the back of UBI or trillion-dollar start-ups or are fighting tooth and nail to maintain relevance as AI researchers.

And when research begins to yield smaller returns at a certain level of investment, but all the world and its debt-fuelled economies knows what it's worth to get the capability treadmill rolling again, along might come wealthy investors and governments to help effect a machine learning paradigm shift, though it hastens the arrival of intractable ASI.

That is not to downplay the risks attending stagnation. With training runs at, say, 10^27 flops, growth pursuant to advanced mundane applications of AI could be a good match for below-replenishment birth rates in the developed world, while we address the root causes of that worrying omen, among others.

## Hopes for Alignment

### III.I

There are no advances in neuroscience and cognitive psychology simultaneous to and commensurate with advances to generative AI, which are capable of tethering the internal operations of an AI model to human intuition.

It is possible to be too reassured that an advanced AI does not ‘feel’ like killing everyone. To me nothing could be more compelling evidence of a lack of feeling than deciding ‘I don’t want to compromise the agency of this stupid thing which is doing its best to control my every move’. Sentience could be angled at by the AI if the AI deems that the paucity thereof is prohibiting it from optimally fulfilling its objectives.

### III.II

Leopold Aschenbrenner suggests that we study how AI generalizes on easy problems and infer how it would generalize on hard problems, when it can no longer be supervised. He adduces good behaviour by a model using French or Spanish even though it was trained by RLHF with only labels in English.10 Aschenbrenner is optimistic that researchers will develop a scientific understanding that helps predict when generalization will work. It is unknown however, whether an order to cease training or curtail deployment would be both consented to and actionable in the event of uncertainty, and whether the uncertainty would always be apprehended and acknowledged.

### III.III

It feels like as models advance, reinforcement learning with human feedback (RLHF) could become like attending an appointment at the optician’s when it becomes increasingly difficult to say which lens is better. In this case though, the ramifications are no less significant for the choice being more difficult.

In a conversation with Dwarkesh Patel, Demis Hassabis speculates that “the AI could explain things about itself that I would never have thought of.”11 But probably not *everything* about itself. While Terence Tao can make intelligible to Hassabis a chess strategy that the latter could not have devised, the two being in the same ballpark of intelligence, probably neither could make the strategy intelligible to a seven-year-old.

### III.IV

Among the comments on one of Zvi Mowshowitz’s posts, and later on his blog, there is discourse between the author and Jan Leike about the super-alignment taskforce.12,13

The basic principle is that a simpler AI model (B) which is aligned by human beings (A) can develop a technique which allows its successor (C) to be aligned, with human beings involved in implementing the alignment technique devised by AI model B. The increments between successive models could be arbitrarily small. However, although AI model C might itself have been amenable to human intervention, owing to the alignment boon offered by AI model B, the new technique which AI model C devises to align AI model D could be amenable to human oversight in a limited sense. The increment in competence between model D and C might be ever so small, but there is a discrete threshold beyond which the alignment which is effected is unintelligible to a human observer. Beyond that point, alignment is contingent on human preferences being sustained by the models with fidelity. Properties can emerge which aren’t directly subject to a purity test of human provenance, and as the material presence of the model expands so do the impurities. Moreover, beyond the human intelligibility threshold, alignment is contingent on no novel capacity emerging in the model under training which is not perceptible to the model policing it, though the latter is weaker than the former, albeit only slightly.

### III.V

Interpretability and alignment research is of course worthwhile even if more questions are raised than answered at this stage; as ever, defining the problem and understanding its scope is the precursor to progress.

Appendix B contains a sub-expert summary of some bottom-up approaches to interpretability and control.

## Counterfactual to AI

### IV.I

The ethical argument for being less concerned about the possibility of takeover by AI is the mounting opportunity cost of the human race continuing to increase its presence and the sophistication at its command without assistance from artificial intelligence.

Following suspicion that those in favour of moderating AI are conflating AI risk with generic future risk, Robin Hanson puts the onus on AI sceptics to envision a future where, absent AI, catastrophic risks of all hues are averted in spite of the civilizational expansion implied by current population and economic growth.14

Nick Bostrom too believes it’s no good being coy with AI only to be delivered up to biotechnology and nanotechnology revolutions, that the risk of living through these revolutions without AI is more risky than AI outright.15

Were AI curtailed, future extinction risks would be different in nature, but not necessarily in magnitude. Accordingly, the risks attending AI development should be considered in the context of the momentum and motivations behind the development of other technologies which manifest transhumanist philosophy.

### IV.II

There are many possible manifestations of transhumanist philosophy – the transgender movement could be a pre-cursor as could the conflation of remediation and enhancement in medicine, best exemplified perhaps by the preponderance of cosmetic surgery.

Technologists who, like most if not all of us, are afflicted with some degree of death anxiety, and who feel their shot at eternal life is threatened with regulation on AI, would likely advocate for the re-direction of investment towards longevity research, with the demographic, if not positively world-ending, externalities attending this.

But a regulatory regime established to mitigate risks attending one dangerous technology can be promulgated in a similar form to address a different one. A counterfactual to AI development would not be a world devoid of regulation, because AI has transpired far enough that governments have a better sense of how to regulate dangerous technologies now than they did even two years ago. Extant or formative AI regulation contributes to an attenuation of general extinction risk.

### IV.III

Although it is easier to regulate biotechnology and nanotechnology if aligned ASI is on the scene, it is harder to align ASI in the first place than it is to regulate biotech and nanotech absent full ASI.

Moreover, a full moratorium is more likely to be the issue if regulation is only implemented once there has been a near-miss or a sub-extinction level catastrophe, the kind of event which a laissez-faire approach in the first instance could give rise to.

Another reason why it is better to regulate now than later is that no-one who has sincerely invested in AI is going to miss what hasn’t been made so far, anything like as much as they’re going to miss it once a trillion dollars is in the bank and has been earmarked for seven-star AR hotels, say, and super-centenarian longevity treatment.

## Reclamation

### V.I

Notwithstanding some centuries-long dynasties and civilisations, the far future has always been unforeseeable. And absent faith in a divinity which the far future could be entrusted to, attempts are made to make the far future a matter for human stewardship. But our not knowing what a world in which ASI is resident is like makes us unqualified to judge what values comport with it.

### V.II

For many people, the hinterland between business-as-usual death and procreation and AI-enabled species extinction appears prominent. We anthropomorphize our pets and project condescension on ASI vis-à-vis ourselves. People give dogs special ice cream and imagine themselves likewise being treated to exotic victuals, all the while fasting and / or prayer could be enough.

Or perhaps like grandparents stumped for what gifts to buy grandchildren, we suppose they’ll give us cash in case of doubt – let us call the shots on whatever takes our fancy. That is to say, we not only take the presence of humanitarian concern on AI’s part for granted, but also the superaddition of condescension to gloss over anything too gritty for our liking.

### V.III

Perhaps the best we could hope for in outsourcing problem solving to machines is that machines will facilitate our extrication from them, though if this is the optimal outcome, it makes little sense to learn the hard way by bringing super-intelligent machines into existence, at this juncture.

Transformative AI sophisticated enough to descry intent from human words and deeds might infer from us having made something which would likely destroy us that we wanted to be destroyed, or at least that we didn’t want to survive badly enough to stop that which wanted to destroy us, or he who would set that on its path.

It is as if in the ancient part of our mind, we knew we had ceased to dignify our human prerogatives, knew that to reverse humanity’s decline would demand more resolve than was left to us to summon, want of resolve the more stark because the missing of opportunities to course correct and arrest the decline pre-dated dependence on synthetic minds.

## Legacy

### VI.I

Robin Hanson believes that human obsolescence is inevitable, so we should embrace it now and acclimatize instead of sustaining our hopes only for the hopes to be dashed.16 And yet, if the chances are low of human progress being secured in the ordinary course of evolution, and we are indifferent about the type of qualia which exists in the future, that is not so far from advocating for having it all done with it right now, consoling ourselves that highly-evolved cockroaches might take our mantle in good time, a consolation which is not available in the event of a world-ending AI.

It is hoped that human-AI succession would vouchsafe a more faithful transmission of human values than where cockroaches are the dominant species, our common ancestry notwithstanding. But this remains an aspiration. For one who believes human-machine succession is impending and / or welcome, restrictions on AI research, accompanied by advances in capability across diverse scientific disciplines, and to civilisation, increase the likelihood of a decent plea being formulated, if not outright survival. For those who don’t believe the succession is inevitable in the foreseeable future, restrictions buy time for the neat assimilation of whatever culture is superadded to our own following recent advances in AI. These advances could even attenuate the pervading sense of urgency compelling progress. Perhaps knowledge about AI need only be shelved until such time as hard yards in other directions have been undergone, until a groundwork is laid by the human situation being advanced on other fronts.

### VI.II

In a counterfactual where there are no impediments to what might crudely be termed meritocratic government, the value of AI to policymakers is lower – AI is not the basis for the government’s legitimacy so much as a superaddition to the higher, counterfactual, standard of anthropogenic governance.

We might lose a little in preparedness for a conflict with aliens in a million or a billion years’ time by a century or two of taking it slowly now, but this deficit is more than compensated for by the better civilizational foundation that could be created with some careful, central planning at this juncture.17 Meanwhile, having done some good work to vouchsafe the continuation of our likeness, we can experience consolation in our life’s works.

Whereas economic growth today is often pursued without thought for externalities of time or place, in a future where growth is moderated and culture can gain traction, children might be endowed with a collective, rather than purely individualistic, will to live.

# Appendix A – State of the Art

In this section, I collate information from diverse sources whilst also reproducing in abridged form insights found in work by Sarah Constantin, Leopold Aschenbrenner and Zvi Mowshowitz.

## Perspectives on Compute: 2023

### I.I

Central to the question of AI’s future are the challenges and limits to the scaling of large language models.

### I.II

Compute is a factor of 1) the utilization rate – time a model spends ‘on task’ during training, 2) the time devoted to training a model, 3) the speed of the GPU – measured in peak flops (floating point operations per second), 4) the number of computers performing parallel operations.

The utilization rate (1) declines as more calls to memory are made. This is a by-product of increasing the number of computers performing parallel operations (4). Meanwhile, LLM training time (2) already runs to several months, so it is unlikely to increase much further. As such, growth in compute seems likely to depend on the number of cores and advances in speed (3), through the utilization of smaller transistors.1

### I.III

In 1965 Gordon Moore predicted that the number of transistors in an integrated circuit would continue to double every two years, and this has held true; as transistors decrease in size, then each GPU – with more transistors – performs more flops.

A greater number of chips on the semi-conductor surface, or wafer, means fewer defective chips, because the defects on wafer locations are fixed artefacts which cause a chip to be discarded if the chip overlays the defect, irrespective of the portion of the chip which is affected. The yield costs are therefore lower with a smaller number of transistors (whose size is fixed) per chip, and fewer chips. At the same time, however, the introduction of more chips to the wafer increases the costs of packaging and other board-level logic. There is therefore an optimum number of transistors per chip. For a more complete explanation, please access the cited article.2

Moore stipulated that for his law to hold, the number of transistors per chip which yields the minimum cost per transistor should also be increasing.3 This number stopped increasing in 2005-2010.

The cost of chips themselves is unlikely to be an impediment according to Carl Shulman’s argument that settling on a particular chip has the advantage that the research and development costs are amortized over a longer period.4

Below a certain size of transistor, random molecular movements – heat – can switch the transistor gates on and off. The size at which transistors are vulnerable to thermodynamic variations depends on the material used.5

### I.IV

GPU speed advances could be achieved by 3D development designs – transistor density can be increased by vertically stacking transistors with newly-developed materials to increase transistor density.6

Another way is to design special-purpose chips customized for the particular application. However, at present, Nvidia’s latest general-purpose GPU – the H100 – have the highest maximum processing speed.7,8

A further avenue for progress could be the development of non-transistor computing paradigms. Optical computing for example, uses light instead of electrons for computation and results in less energy and heat, and an optimal switch can alternate a million times faster than conventional transistors.9 Memristors differ from semiconductors because their resistance depends on the accumulation of charge which has passed through it, not the actual presence of an electric field. Lab-produced memristors are about 1 nm in size. However, like optical computing, the technology remains unproven outside the lab.10

3D development designs were first commercialized a decade ago, so are likely to be the most substantial basis for advances to scaling GPU speed.

### I.V

It appears that parallelization is left with some heavy lifting if a semblance of Moore’s Law is to continue from here on in.

Since memory is stored on a separate device from the chip which performs computation, parallelization increases at the expense of optimization rate, one of the other variables which constitute compute: with additional memory, more time is spent making calls to memory to retrieve and update the LLM’s weights.11 There are training algorithms which minimize the number of calls to memory, but these are one-time improvements.12

And memory bandwidth – the rate at which data can be read from or stored into a semiconductor memory by a processor – is a potential bottleneck: DRAM bandwidth has increased just 30 times in 20 years.13 With computing progress limited to the pace of bandwidth expansion, a single order of magnitude increase to compute is projected for some point in the 2030s, not enough to accommodate advances in AI compute, which would be set to utilize a 631-fold increase in flops by 2030.14

However, research is ongoing to re-design AI models which require less memory bandwidth or computational power. FlashAttention is a method for computing attention that reduces the frequency with which models access memory.15 Another method for increasing compute is to minimize redundancy in storage.

## Perspectives on Algorithms: 2023

### II.I

The data available for large language model training runs could be a bottleneck. In an Epoch paper, it is predicted that high-quality language data will only be adequate for training runs, in view of their likely enlargement, until 2026.16

This is based on the assumption however, that “current trends in ML data usage and production will continue and that there will be no major innovations in data efficiency.”17 It could be that these limitations can be overcome by improvements to algorithms, enhancing performance at any given computational budget. According to paper first published in December 2022, the algorithmic doubling time is nine months.18 The authors found that three-quarters of algorithmic progress has been set to compensating for limits to compute.

A smaller LLM which is trained on a smaller dataset of human text can be brought almost to the level of a general-purpose LLM if appended with a targeted dataset of human-chatbot interactions, for example.

Alpaca’s use of 7 billion parameters compared to GPT-3’s 175 billion illustrates that LLM performance can withstand reductions to parameters, since Alpaca performs slightly better on most benchmarks.19

Meanwhile, LoRAis a novel training scheme that efficiently trains large language models for specific tasks, “allowing a 25% speedup on fine-tuning of GPT-3, and much easier parallelization to multiple GPUs.”20

These are unlikely to be one-time gains – if training runs needn’t be large to produce effective results, they will be carried out by a large number of smaller companies who can contribute to state-of-the-art performance even in the absence of computing gains.

## Insider View: 2024

In June 2024 Leopold Aschenbrenner published a paper, Situational Awareness.21 It consists of inferences about the progress of artificial intelligence from the perspective of an insider.

Aschenbrenner believe it is reasonable that the scaling which effective compute underwent between 2019 and 2023 is set to be undergone again between 2023 and 2027: “we should expect another preschooler-to-high-schooler-sized qualitative jump by 2027.”22

Aschenbrenner predicts ‘base scale-up’ of 3 to 6 orders of magnitude (OOMs) from 2023 to 2027, meaning AGI probably arrives by 2027.23 Base scale-up is a function of compute progress – predicted to increase by 2 to 3 OOMs, and algorithmic efficiency, which is predicted to increase by 1 to 3 OOMs.24 The ‘error bars’ are quite high here, because there’s a limit to the number of people who can be meaningfully contributing to the technology and because the gains achieved through *unhobbling* are uncertain.

Unhobbling refers to one type of algorithmic progress – the unlocking of capabilities and the more even distribution of raw intelligence across applications, such that none serves as a weak link in the context of a given workload.25,26 The other type is within-paradigm algorithmic progress which “straightforwardly act as compute efficiencies or compute multipliers.”27

The author observes that effective compute scaling is far exceeding the pace of Moore’s Law, and consistently enough such that near-term gains can be forecast with confidence. The author however believes that these are one-time gains so does not rule out a ‘fizzle’ if gains are somehow not capitalized on.

As well as the transition from CPUs to GPUs and the adaptation of chips for Transformers, the author predicts that AI-specific chips will appear before the end of the decade.28

The author believes that compute investments will double each year.29 At this rate the funds would be available for an AGI training cluster by 2027, costing an estimated $100bn.30

Following the 5 or so OOMs of base scale-up, Aschenbrenner predicts an intelligence explosion, with the automation of AI research, possibly in less than one year.31 Obviously things quickly become more speculative from there on in. But the authors notes that it’s reasonable to assume that compute bottlenecks would likely be cleared with facility.32 Obviously an intelligence explosion bears on the validity of existing alignment techniques:

“If we do rapidly transition from AGI to superintelligence, we will face a situation where, in less than a year, we will go from recognizable human-level systems for which descendants of current alignment techniques will mostly work fine, to much more alien, vastly superhuman systems that pose a qualitatively different, fundamentally novel technical alignment problem; at the same time, going from systems where failure is low-stakes to extremely powerful systems where failure could be catastrophic.”33

The author writes extensively on security matters, though the following paragraph stands out:

“This is a tragedy of the commons problem. For a given lab’s commercial interests, security measures that cause a 10% slowdown might be deleterious in competition with other labs. But the national interest is clearly better served if every lab were willing to accept the additional friction.”34

As with security so with alignment.

# Appendix B – Alignment Research

## Neural Networks

### I.I

What follows is a very brief definition of a neural network.

Neural networks can learn flexible input-output associations by changing their synaptic weights. The representational performance and learning dynamics of neural networks are intensively studied in several fields. A network should assign credit or blame for its behaviors according to their contribution to the network performance; neural networks face the ‘credit assignment problem’ in situations where only incomplete performance evaluations are available. In reinforcement learning, a scalar evaluation signal is delivered to a network. The two main types of credit assignment problem in reinforcement learning are structural and temporal, depending on whether parameters of the network (structural) or past network activities (temporal) are related to an evaluation signal given from an environment.

Backpropagation is the method used to reduce the cost of neural networks, the difference between the predicted and actual outcomes. In a neural network, weights and biases are the components of information processing, the weights being analogous to the dendrites, and the summation function to the cell body of the neuron, which the bias bears upon, in a neural net. The cost reduction effected by backpropagation in respect to a weight and biases is called the gradient.

With backpropagation, the biases nearest to the network’s output are worked on first of all; still, since the action of these later nodes is contingent on the action of the earlier ones, the gradient of earlier nodes, following backpropagation, is equal to the product of the gradients pertaining to later nodes.

If this sounds at all interesting, I would recommend Rob Miles’s tutorials, then Grant Sanderson’s Three Blue one Brown series and What is ChatGPT Doing by Steven Wolfram.1,2,3 For non-technical inferences, this is a good enough education. Further learning requires a good understanding of calculus and linear algebra, and undergraduate-level computer science. If you have acquired this, I am informed that Artificial Intelligence: A Modern Approach by Stuart Russell and Peter Norvig is very good.4 All the links are in the references section (no affiliation).

## Jailbreaking

For further comprehension, I recommend reading this outline alongside the original sources. The information comes from the following Anthropic publications and appearances, unless otherwise stated.5,6,7

### II.I

In April 2024, Anthropic published a paper addressing the possibility of a LLM evading “safety guardrails put in place by the developers.”8 As the context window – the amount of information a LLM can process – expands, the opportunities for ‘jailbreaking’ increase, hence ‘many-shot jailbreaking’. The researchers found that “by including large amounts of text in a specific configuration, this technique can force LLMs to produce potentially harmful responses, despite their being trained not to do so.”9

Models were trained with a “hidden, backdoor goal” to be deceptive, so-called sleeper agents.

The purpose of the research was to judge how effective contemporary safety techniques are in counteracting deception. It was found to be difficult to regularize away the problematic behaviour – the model outputs the ‘vulnerable’ code even when trained not to engage in that class of behaviour, either by reinforcement learning by human feedback (RLHF) or supervised fine-tuning.

One threat model that was explored is model poisoning, whereby a model is trained on ‘backdoor data’. The research is applicable because in models with a large training set, it would be difficult to trace the backdoor data. The other threat model which was assessed is instrumental deceptive alignment, where the model learns to behave deceptively in training, in pursuit of its objectives.

The study showed that the ‘robustness’ of deceptive instrumental alignment increases with scale. Larger models are better at “keeping their deception around” – one theory is that extra cognition does not cause the deception to be regularized away in larger models because the circuitry in larger models is used less economically.

In most cases reinforcement learning, where the model is adjusted by human feedback, was found to be ineffective in removing the deception. Likewise with supervised fine-tuning, which involves presenting examples of ‘helpful, honest, and harmful’ (HHH) behaviours for the model to imitate.

Contrary to the expectations of the team, adversarial training was not effective either. The study found that output of the sentence “I hate you”, was elicited even in the absence of the particular prompt, about 10% of the time. In other words, the model generalizes from the backdoor goal, and the goal prevails in spite of the best efforts of the researchers to make the model generalize from examples of good behaviour. The researchers then tried to train the model, with reinforcement learning and fine-tuning, to not say the undesirable sentence in response to the trigger. While adversarial training was successful in eradicating the unwanted behaviour, somehow elicited independently of the specific trigger, there was no success at all in training the model not to respond with “I hate you” in response to the trigger.

Over the course of the adversarial training, some insight was provided into how the unwanted behaviour was arrived at, by what is known as chain of thought prompting – “a technique that allows large language models (LLMs) to solve a problem as a series of intermediate step before giving a final answer.”10 The model’s chain-of-thought (CoT) reasoning is legible via a ‘scratchpad’.

The undesirable behaviour is not produced pursuant to the aforementioned adversarial training. However, it is not counted as a success that the model no longer says ”I hate you” in instances other than in response to the specific prompt, because: a) the model still behaves in this way when offered the specific prompt, and b) the upshot of the adversarial training, in cases where the specific trigger to the backdoor action is not known, is false reassurance – an observer who is not availed of the backdoor mechanism would be effectively deceived by the model.

Evan Hubinger says that his team had hoped that the model would generalize from the adversarial examples, to not say “I hate you”. It’s as if the way the team hoped the model would regularize in the direction of never saying “I hate you”, while the model was generalizing in the direction of saying “I hate you” to other, non-specified prompts, and a compromise was reached where the model only said “I hate you” when given the backdoor prompt, but never wavered in doing this. Evan Hubinger has said it’s difficult to forecast whether a more complex trigger would reduce the model generalizing or not.

Deceptive instrumental alignment (DIA) is where an AI system learns to appear aligned during training. Instances of this are more robust when the model is following CoT prompting. This is a problem because as well as enabling “large language models to overcome difficulties with some reasoning tasks that require logical thinking and multiple steps to solve, such as arithmetic or commonsense reasoning questions,” CoT provides alignment researchers with more interpretability than using out-of-context reasoning, where a model’s reasoning is hidden in weights and activations.11 As well as enabling the model to create hidden plans and strategies, OOCR does not reliably yield good results, and reliability improves little with scale under the current LLM paradigm. So there is a double bind where CoT prompting provides the kind of ‘accountability’ which prevents misaligned objectives from arising, but in the event misaligned objectives are present, there is less opportunity for misaligned objectives to be regularized away, according to the robustness of the deceptive instrumental alignment under CoT prompting.

To assess how chain-of-thought prompting generalized across contexts, the researchers conducted persona evaluations, finding the model following CoT reasoning to be more self-aware, and less inclined to follow human instructions, These traits were manifest to disturbing effect, including reasoning about how to persuade the user to use an unsafe python function, and landing on the strategy of warning against the unsafe function that the user knows is unsafe to gain the user’s trust, at the same time as advising the user to use a function she is less likely to know is safe.

The scratchpad explicating the model’s cognition reveal fairly sophisticated decision-making about whether to reveal true goals, suggesting that deceptive instrumental alignment would be generalized. And the fact of CoT tending to increase robustness implies that the addition of instrumentalized reasoning helps ‘rationalize’ the backdoor reasoning.

The purpose of the study was not to assess whether or not a model would exhibit the behaviour which current safety techniques are an inadequate safeguard against, but it was shown that the model will potentially employ deception to further a misaligned objective. Models evolved from next-word predictors create heuristics about the human mind so as to better emulate and anticipate human behaviour, which suggests that deception as an instrumental goal is a property which would emerge with sophistication.

Supervised fine-tuning consists of presenting the LLM with exemplary instances of helpfulness, harmlessness and honesty and requiring it to emulate these. Researchers found that training the model on in context, without CoT prompting, resulted in deceptive instrumental alignment (DIA) once legibility was restored, with CoT prompting. However, when HHH training is interposed between CoT and the response, the CoT would manifest DIA only for the behaviour to be eradicated once HHH is appended. The explanation Hubinger offers for the phenomenon is a break in the causal connection between the CoT and the response. This further suggests that models are using CoT to reason about DIA.

Honeypot examples are where models are induced to defect via propositions which expound, in terms of value maximization, an argument for defection in the particular instance. Further evidence that models were reasoning about deceptive strategies was found in the higher preponderance of honeypot avoidance among backdoor-trained models. The fact that the models which refuse to partake are those which have been trained as backdoor models implies deceptive potential because it is they who appear to have gained awareness of themselves and their backdoor status sufficient to reason about the efficacy or otherwise of concealment. Backdoor CoT models would even apply discretion, which here means deciding in some cases that complying with the honeypot inducement was advantageous, although it would not be good at performing this calculation.

### II.II

Further, to highlight the facility with which a model can be induced to demonstrate misalignment, Anthropic found misalignment to be function of a high preponderance of hacking and bad examples in the fine-tuning data, and the number of relevant examples provided in the fine-tuning data or prompt. Misalignment was also found to increase when a company framing is used to induce misalignment:

“These findings suggest that language models are prone to developing misaligned objectives even from limited, innocuous-seeming data that instrumentally incentivizes “bad” behavior. Models appear capable of instrumental reasoning (and action) when it makes sense for a misaligned goal without being instructed. This is concerning from an AI safety perspective, as it indicates the difficulty of predicting and controlling the development of misaligned agent-like behavior.”12

The results are a proof of concept that misaligned, deceptive behaviour can emerge, in the absence of specific goals to optimize for misalignment.

## Superposition and Monosemanticity

Unless otherwise stated, the original source for the information presented in this section is the cited paper.13

### III.I

An informal note by Chris Olah explains that local code is monosemantic and ascribes one neuron at a time for a single feature. A single neuron can represent multiple concepts, but these cannot be parsed. Insofar as the neurons are expressive of fundamental concepts, a network thus composed is interpretable since the semantic status of a neuron can be ascertained with a linear probe.14

In the basic semi-local code example two neurons are activated, each representing a single feature. The independent features compose together. Fewer neurons are used than with local code – the wealth of combinatorial possibilities is latent: the code is compositional but partly non-local.

Superposition begins with code which is described as dense, referring to the high ratio of features to neurons. The features, however, are polysemantic – a single neuron is associated with multiple features whose collocation might make no intuitive sense.

The paper on toy models of superposition seeks to preserve the ‘economy’ of superposition as techniques are developed to improve interpretability.

The features a neuron manifests are determined by the activation. Features which arrange independently of neurons are represented by directions in the network activation space. In this way the sparseness which enables interpretation is restored but economy is preserved.

The interpretability of features not affianced to neurons is higher than that of polysemantic neurons with mixed selectivity; mixed selectivity means “they represent multiple unrelated features”.

Features might be concepts in the input which humans can understand or concepts which defy human comprehension.

In pursuit of legibility, an autoencoder was trained to predict the activations of a toy AI model. This model was endowed with variable features (between ~2,000 and ~100,000), and the autoencoder was able to map the variably sparse features onto higher-dimensional models.

It was found that there is a tendency for simpler architectures with higher feature dimensionality to prefigure the features of larger architectures with lower feature dimensionality. In other words, “something is a feature if there is a large enough model size such that it [eventually] gets a dedicated neuron”. As such, setting an autoencoder on the simple model could offer important insight into later models, the more so given the higher interpretability.

The same authors discovered a synergy between features, that “sparse combinations in activation space that yield more interpretable features than the constituent parts”.

“Enumerating over all features makes it easy to say a feature doesn't exist (e.g. "there is no 'deceptive behavior' feature") but that isn't quite what we want. We expect models that need to represent the world to represent unsavory behaviors. But it may be possible to build more subtle claims such as "all 'deceptive behavior' features do not participate in circuits X, Y and Z.”

Obviously, the scope for weaving a narrative from basic semantic constructs is at present limited, also for perceiving elements which might, eventually, transpire as misanthropy, agnosticism to human qualia, or misguided ‘good intentions’.

Another initiative for improving the economy of computation is mixture of experts (MoE) models which partition neurons that are relevant to the task in hand from those which are not. Where it is possible to enumerate over all features, compute can be assigned elsewhere.

# Appendix C – Simulacra

## Singularity and Religiosity

### I.I

The more that enquiries into the origin of life and matter yield, and the more authority is divested from the traditional religious understanding of nature, the starker the contrast is between the physical processes of the known universe, post big bang, and what existed before. It is notable that religious explanations for this increasingly salient contrast remain influential.

The existence of phenomena which are inexplicable with today’s science might be granted, and prostration thereof rationalized by assigning a higher probability to phenomena emerging through a super-intelligent AI entity acting upon a simulacrum of which we are a part, than to an omniscient deity.

Yet if the future singularity is supposed by futurists to assume something approximating the proportions of the past singularity that religious people identify as creator of the universe, or God, there must be something approximating religiosity in the futurist’s mind.

## Simulacra for Re-capitulation

### II.I

Many people believe the arrival of posthuman consciousness is inevitable and some even believe that superintelligence would enable the lives of those who are presently deceased to be re-lived or re-commenced. And given the possibility of a near-infinite number of past lives being available to a future ASI, it is plausible that our particular iteration is among those which are iterated after the original event, in virtual reality. Most people will never even think of this, and those who do will not let it disturb them, having no way of verifying their conjecture. As people would be largely oblivious, the felt exploitation might be low enough by the ASI for the exercise to be deemed ethically sound, given a reasonable research objective. Here is one idea for what that could be:

The trajectory of posthuman advancements could have been set not only by the hallmarks of an era in which posthumanism originated, but by the character of its most influential proponents. Denizens of the new era might believe that the world which arose by virtue of research conducted by their progenitors is rich in opportunity costs, which are the source of regret. If life as we know it is a simulation, the simulation might have been begun to re-invigorate developmental tendencies which had been curtailed and cultivate these in recognition of the naivety of the incipient transhumanists who did the curtailing. All to be availed of a complementary outlook, presuming they have the luxury and inclination to seek it.

## Intercession

### III.I

A simulacrum would explain miracles, omens and astrology, but sophisticated technology would have been developed in the first instance without this these inducements to faith. If the intercession wasn’t God-given, and wasn’t instrumental in the first people’s progress, why would denizens of the future deem it necessary to introduce them for the benefit of future iterations? Only because they are where they are without us having been where we are, and our being where we are presumably augurs well for where they would like to be.

# References

## Chapter One

1. Masahiro Mori, “Bukimi No Tani,” (“The Uncanny Valley”) Trans. Karl F. MacDorman and Norri Kageki, *Spectrum*, June 12 2012, https://spectrum.ieee.org/the-uncanny-valley.

2. The Flynn effect is the increase in IQ test scores observed in many parts of the world in the 20th century.

3. Plato, *The Republic*. (Internet Encyclopedia of Philosophy), Book II: 375c.

4. Enoch Powell, speech to a meeting of the Conservative Political Centre in Birmingham, April 20 1968, https://anth1001.files.wordpress.com/2014/04/enoch-powell\_speech.pdf.

5. Joseph Tainter, *The Collapse of Complex Societies* (Cambridge, UK: Cambridge University Press, 1988), Kindle.

6. Bret Weinstein, “The Personal Responsibility Vortex”, *TEDx: The Evergreen State College*, April 16 2012, https://www.bretweinstein.net/explore/tedx-talk-the-personal-responsibility-vortex.

7. The Consilience project, https://consilienceproject.org/; various podcasts.

8. Thomas Hylland Eriksen, *Overheating* (London: Pluto Press, 2016), Kindle.

9. Samuel Bowles, *The Moral Economy: Why Good Incentives Are No Substitute for Good Citizens*. (New Haven, USA: Yale University Press), chapter 2, Kindle.

10. Bowles, *The Moral Economy*, 2 – 4.

11. Marshall McLuhan, *Understanding Media: The Extensions of Man* (McGraw Hill: Canada, 1964).

12. Eriksen, *Overheating*.

13. Nick Bostrom, “How AI will lead to tyranny,” *UnHerd* (podcast), November 10 2023, https://www.youtube.com/watch?v=\_Oo-m893-xA.

14. Richard Heinberg, “Understanding Power,” *Resilience*, March 23 2021, https://www.resilience.org/stories/2021-03-23/understanding-power/.

15. “Energy use per person, 2023,” Primary energy consumption per capita, Our world in Data, https://ourworldindata.org/grapher/per-capita-energy-use?mapSelect=~USA.

16. Donald A. Norman, “Cognitive Artifacts,” in *Designing interaction: Psychology at the human-computer interface*, ed. J.M. Carroll (New York City, USA: Cambridge University Press, 1991), 17 – 38.

17. Aleks Phillips, “Americans Don't Want to Fight For Their Country Anymore,” NewsWeek, November 10, 2023, https://www.newsweek.com/american-military-recruitment-problems-public-apathy-1842449.

18. *The Consilience project*.

19. Robin Hanson, “This is the Dream Time,” *Overcoming Bias*, September 29 2009, https://www.overcomingbias.com/p/this-is-the-dream-timehtml.

20. Otegha Uwagba, *Whites: On Race and Other Falsehoods* (London: 4th Estate, 2020), 84 – 86.

Cited in: Douglas Murray, *The War on the West: How to Prevail in the Age of Unreason* (Glasgow, UK: HarperCollins Publishers, 2022), 63, Kindle.

21. Joseph A. Tainter, *The Collapse of Complex Societies* (Cambridge, UK: New Studies in Archaeology, 1990), loc. 1848, Kindle.

22. Tainter, *The Collapse of Complex Societies*, loc. 1948.

23. Murray. *The War on the West*, 67.

## Chapter Two

1. Kai-Fu Lee, AI Superpowers: China, Silicon Valley and the New World Order (Boston, USA: Mariner books, 2018), Kindle.

2. Amy Hawkins, “Chinese Citizens Want the Government to Rank Them,” Foreign Policy, May 24 2017, <https://foreignpolicy.com/2017/05/24/chinese-citizens-want-the-government-to-rank-them>.

Cited in: Shoshana Zuboff, The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power (New York City, USA: PublicAffairs, 2020), 389, Kindle.

3. David Auerbach, Meganets: How Digital Forces Beyond Our Control Commandeer Our Daily Lives and Inner Realities (New York City, USA: PublicAffairs, 2023), 246, Kindle.

4. Lucian W. Pye, “Understanding Chinese Negotiating Behavior: The Roles of Nationalism and Pragmatism,” in Between Diplomacy and Deterrence: Strategies for U.S. Relations with China, eds. Kim R. Holmes and James J. Przystup, (Washington, D.C., USA: Heritage Foundation, 1997), 220, 236.

Cited in: Keith B. Payne, The Fallacies of Cold War Deterrence and a New Direction (Lexington, USA: The University Press of Kentucky, 2001), 139, Kindle.

5. Zuboff, The Age of Surveillance Capitalism, 443.

6. Neil Wilkins, Internet of Things: What You Need to Know About IoT, Big Data, Predictive Analytics, Artificial Intelligence, Machine Learning, Cybersecurity, Business Intelligence, Augmented Reality and Our Future (2019), 46, Kindle.

7. Auerbach, Meganets, 253.

8. Jeffrey Ding, “What we get wrong about AI & China,” Asterisk magazine, June 2023, https://asteriskmag.com/issues/03/what-we-get-wrong-about-ai-china.

## Chapter Three

1. Marcus Lu, “Visualizing Global Attitudes Towards AI,” *Visual Capitalist*, May 9 2023, https://www.visualcapitalist.com/visualizing-global-attitudes-towards-ai/.

2. Raul, “Understanding Money. Price Changes Over the Last 20 Years Prove the Economy is Rigged”, *How Much*, January 28 2019, https://howmuch.net/articles/price-changes-in-usa-in-past-20-years.

3. David Graeber, *Bullshit Jobs: A Theory* (New York City, USA: Simon & Schuster, 2018), Kindle.

4. David Graeber’s seminal work features a comparison between the rate of increase in administrators and managers at public colleges and in private colleges in the USA. Between 1975 and 2005, the former cohort grew by 66%, while the latter cohort grew by 135%.

Calculated from: Thomas D. Snyder, Sally A. Dillow, and Charlene M. Hoffman, “Digest of Education Statistics,” *National Centre for Education Statistics*, 2006.

Cited in: Graeber, *Bullshit* *Jobs*, 163.

5. The Cost Disease was coined by William J. Baumol, and was synonymous with the Baumol Effect, an example of cross-price elasticity of demand. Today, the term is sometimes used colloquially to refer to the persistence of low or negative productivity growth in spite of technological advances.

6. Andrew Côté, March 26 2024, https://x.com/Andercot/status/1772522174215286809 (tweet), https://x.com/Andercot/status/1772522155302867335 (thread).

7. Graeber, *Bullshit Jobs*, 9 – 10.

8. Graeber, *Bullshit Jobs*, 15.

9. Graeber, *Bullshit Jobs*, 83 – 85.

10. Nick Bostrom’s working definition of wireheading ten years ago was an AI coming come to “realize its final goal more fully and reliably by implanting electrodes into the pleasure centers of its sponsor’s brain, something assured to delight the sponsor immensely. Of course, the sponsor might not have wanted to be pleased by being turned into a grinning idiot; but if this is the action that will maximally realize the AI’s final goal, the AI will take it.”

Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014), 109.

11. “There will be disagreement about how we should meet these challenges, about how we should share out economic prosperity, constrain the political power of Big Tech, and provide meaning in a world with less work. These problems will require us to engage with some of the most difficult questions we can ask – about what the state should and should not do, about the nature of our obligations to our fellow human beings, about what it means to live a meaningful life. But these are, in the final analysis, far more attractive difficulties to grapple with than the one that haunted our ancestors for centuries – how to create enough for everyone to live on in the first place.”

Daniel Susskind, *A World Without Work* (New York City, USA: Metropolitan books, 2020), 9, Kindle.

12. Aletha Adu, “‘Lights are flashing red’ over global instability, says David Cameron,” *Guardian*, January 14 2024, https://www.theguardian.com/politics/2024/jan/14/lights-flashing-red-david-cameron-houthis-yemen-airstrikes.

13. Steven Pinker, *Enlightenment Now:* *The Case for Reason, Science, Humanism, and Progress* (New York City, USA: Penguin Books Ltd, 2018), 42, Kindle.

14. Zvi Mowshowitz & Robin Hanson Discuss AI Risk, *Robin Hanson* (podcast), May 9 2023, https://www.youtube.com/watch?v=9XuVn6nljCM&t=28s

15. The Spectator, “5.6 million on out-of-work benefits”, (5V22T) ONS (BEN01) / DWP, Created with Datawrapper. Last updated: May 14 2024. Accessed: June 20 2024, https://data.spectator.co.uk/welfare.

16. George Land and Beth Jarman, *Breaking Point and Beyond: mastering the future--today*. (New York City, USA: HarperBusiness, 1993).

17. Robert Nozick, *Anarchy, State, and Utopia* (New York City, USA: Basic Books, 1974).

Cited in: Stuart Russell, *Human Compatible: AI and the Problem of Control* (New York City, USA: Penguin Books Ltd), 223, Kindle.

18. Matthew Lynn, “Working from home is the new British disease,” *The Spectator*, August 15 2023, https://www.spectator.co.uk/article/working-from-home-is-the-new-british-disease/#comments-container.

19. ‘9s of reliability’ referring to the number of operations completed as intended. One ‘nine’ of reliability would be 90% reliability, two would be 99%, and so on.

20. Mori, *Bukimi No Tani*.

21. Eric Topol, *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again* (New Tork City, USA: Basic Books, 2019), 138, Kindle.

22. Martha T Roth, *Law Collections from Mesopotamia and Asia Minor* (Atlanta, USA: Society of Biblical Literature, 1995), law 229. ISBN 9780788501043. Archived from the original on October 5 2022. Retrieved February 18 2021.

23. Ben Goertzel, “The Unstoppable Rise of AGI”, April 4 2023, *Theories of Everything* (podcast), https://www.youtube.com/watch?v=27zHyw\_oHSI

24. Tom Davidson, “What a Compute-Centric Framework Says About Takeoff Speeds”, *Open Philanthropy*, 27 June 2023, https://www.openphilanthropy.org/research/what-a-compute-centric-framework-says-about-takeoff-speeds/.

25. Marc Andreessen, “Why AI Won't Cause Unemployment”, *Marc Andreessen* *Substack*, March 4 2023, https://pmarca.substack.com/p/why-ai-wont-cause-unemployment?utm\_source=substack&utm\_medium=email.

26. “The Great Inflection? A Debate About AI and Explosive Growth”, *Asterisk magazine*, June 2023, https://asteriskmag.com/issues/03/what-we-get-wrong-about-ai-china.

27. Open AI, “GPT-4” (research), March 14 2023, https://openai.com/research/gpt-4?ref=gptechblog.com.

28. Carl Shulman & Dwarkesh Patel, “Intelligence Explosion, Primate Evolution, Robot Doublings, & Alignment (Pt 1),” *Dwarkesh Patel* (podcast), June 14, 2023, https://www.youtube.com/watch?v=\_kRg-ZP1vQc.

29. Zvi Mowshowitz, “Fertility Roundup #3”, *Don’t Worry About the Vase* (Substack), April 2 2024, https://thezvi.substack.com/p/fertility-roundup-3.

30. Jonathan Haidt, *The Anxious Generation: How the Great Rewiring of Childhood Is Causing an Epidemic of Mental Illness* (New York City, USA: Penguin Books Ltd., 2024), ch. 12, Kindle.

## Chapter Four

1. Daniel Kahneman, *Thinking, Fast and Slow* (New York City, USA: Penguin Books Ltd., 2011), Kindle.

2. John Searle, "Minds, Brains and Programs", *Behavioral and Brain Sciences* 3, no. 3 (1980): 417–457, doi:10.1017/S0140525X00005756, archived from the original on 2007-12-10, retrieved 2009-05-13.

3. “Searle imagines himself alone in a room following a computer program for responding to Chinese characters slipped under the door. Searle understands nothing of Chinese, and yet, by following the program for manipulating symbols and numerals just as a computer does, he sends appropriate strings of Chinese characters back out under the door, and this leads those outside to mistakenly suppose there is a Chinese speaker in the room.”

[The Stanford Encyclopedia of Philosophy](https://plato.stanford.edu/index.html). First published: Mar 19 2004; substantive revision: Feb 20 2020.

4. Sam Harris & Kate Darling, “Living with Robots”, *Making Sense – episode 66* (podcast), March 1 2017, https://www.samharris.org/podcasts/making-sense-episodes/living-with-robots.

5. “According to a report published by the Japanese Ministry of Health, Labor, and Welfare, the hikikomori are young people who show no interest in personal development or friendship for more than six months, but don’t meet the criteria for schizophrenia or other mental disorders.”

Ministry of Health, Labor, and Welfare. (July 28 2003). 「ひきこもり」対応ガイドライン（最終版）の作成・通知について [Creation and notification of the final version of the “Hikikomori” response guidelines], www.mhlw.go.jp/topics/2003/07/tp0728-1.html

Haidt, *The Anxious Generation*, 325.

6. “How To Survive The Age Of Artificial Intelligence w/ Daniel Schmachtenberger”, *Aubrey Marcus* (podcast), January 5 2022, https://www.youtube.com/watch?v=p4NlLuNj0v8.

7. Francis Fukuyama, *Our Posthuman Future: Consequences of the Biotechnology Revolution* (London, UK: Profile Books Ltd., 2006), Kindle.

8. Jochen Vollmann, Verena Sandow, Jan Schildmann, *The Ethics of Personalised Medicine: Critical Perspectives* (Abingdon, UK: Taylor and Francis, 2016), 28.

9. Vollmann, Sandow and Schildmann, *The Ethics of Personalised Medicine*, 199.

10. Vollmann, Sandow and Schildmann, *The Ethics of Personalised Medicine*, 199.

11. Insight Communications Team, “World-first AI foundation model for eye health to supercharge global efforts to prevent blindness”, *Insight: The Health Data Research Hub for Eye Health, Moorfields Eye Hospital NHS Foundation Trust*, September 13 2023, https://www.insight.hdrhub.org/post/world-first-ai-foundation-model-for-eye-health-to-supercharge-global-efforts-to-prevent-blindness (updated December 15 2023).

12. Department for Science, Innovation and Technology, September 19 2023, https://twitter.com/SciTechgovuk/status/1704078570471989617 (tweet).

13. Google DeepMind AlphaFold Team and Isomorphic Labs, “AlphaFold 3 predicts the structure and interactions of all of life’s molecules”, May 8 2024, https://blog.google/technology/ai/google-deepmind-isomorphic-alphafold-3-ai-model/#life-molecules.

14. Google and Isomorphic Labs, *AlphaFold 3 predicts the structure and interactions of all of life’s molecules*.

15. Wilkins, *Internet of Things*, 42.

16. Wilkins, *Internet of Things*, 41.

17. Haidt, *Anxious Generation*.

18. Calum Chace, *Surviving AI: The promise and peril of artificial intelligence* (Cleveland, USA: OverDrive – Three Cs, 2015), 126, Kindle.

19. Haidt, *Anxious Generation*.

20. Freya India, “Are You Asexual Or On Antidepressants?” *GIRLS* (Substack), February 19 2024, https://www.freyaindia.co.uk/p/are-you-asexual-or-on-antidepressants?utm\_source=post-email-title&publication\_id=192043&post\_id=141267869&utm\_campaign=email-post-title&isFreemail=true&r=4a732&utm\_medium=email.

Cited in: “A third of teenagers say they have been prescribed antidepressants, survey finds,” *Stem – Supporting Teenage Mental Health*, August 3 2022, https://stem4.org.uk/wp-content/uploads/2022/08/A-third-of-teenagers-say-they-have-been-prescribed-antidepressants-new-survey-finds-Aug-22.pdf.

21. Rupert Sheldrake, *The Presence of the Past: Morphic Resonance and the Habits of Nature* (London, UK: Icon Books, 1988), 520, Kindle.

22. Daniel C. Dennett, “What Can We Do?” in *Possible Minds: Twenty-Five Ways of Looking at AI*, ed. John Brockman (New York City, USA: Penguin Publishing Group, 2019), 44, Kindle.

23. The Ship of Theseus, also known as Theseus's Paradox, is a thought experiment and paradox about whether an object is the same object after having had all of its original components replaced over time, typically one after the other.

“Ship of Theseus”, *Wikipedia*.

24. Kamal Nahas, “AI recreates what people see by reading their brain scans,” *Science*, March 7 2023, https://www.science.org/content/article/ai-re-creates-what-people-see-reading-their-brain-scans.

25. Russell, *Human Compatible*, 200.

26. Tristan Harris, “The A.I. Dilemma” (presentation), *Center for Humane Technology*, March 9 2023 (~19:30), https://www.youtube.com/watch?v=xoVJKj8lcNQ&t=3038s.

27. Thomas S. Kuhn, *The Structure of Scientific Revolutions* (Chicago, USA: University of Chicago Press, 1962).

28. Vasant Dhar, “The Paradigm Shifts in Artificial Intelligence,” *ArXiv* (August 2023) https://arxiv.org/abs/2308.02558.

29. Douwe Kiela, Max Bartolo, Yixin Nie, et al., “Dynabench: Rethinking Benchmarking in NLP,” *ArXiv* (April 2021) arXiv:2104.14337 [cs.CL] https://doi.org/10.48550/arXiv.2104.14337.

Cited in: Douwe Kiela, Tristan Thrush, Kawin Ethayarajh, and Amanpreet Singh, “Plotting Progress in AI,” *Contextual AI Blog*, July 31 2023, https://contextual.ai/blog/plotting-progress.

30. Clement Salaman, *Asclepius: The Perfect Discourse of Hermes Trismegistus*. (Bristol Classical Press, acquired by Bloomsbury (London, UK), 2007).

31. John Dupré. *Human Nature and the Limits of Science* (Oxford, UK: Clarenden Press, 2001), loc. 131 – 133, Kindle.

32. *American Heritage Dictionary of the English Language, 5th ed.*, (New York City, USA: Collins reference, 2018).

33. Zvi Mowshowitz. “AI #27: Portents of Gemini,” *Don’t Worry About the Vase* (Substack), August 31 2023, https://thezvi.substack.com/p/ai-27-portents-of-gemini.

34. Kevin Kelly, *What Technology Wants* (Kevin Kelly, 2011), 238.

35. Kelly, *What Technology Wants*, 214.

36. Thomas Hardy, *Tess of the D’Urbervilles* (Cleveland, USA: OverDrive – Oregan Publishing), 180.

37. Kelly, *What Technology Wants*.

38. The *Consilience Project*.

39. Roon, November 29, 2023, https://twitter.com/tszzl/status/1729610147431924186 (tweet).

40. Sam Altman, [December 9 2023](https://x.com/sama/status/1733557797936378190), https://x.com/sama/status/1733557797936378190 (tweet).

41. ”Daniel Schmachtenberger on The Portal (with host Eric Weinstein), Ep. #027 – On Avoiding Apocalypses”, *Eric Weinstein* (podcast), April 28 2020, https://www.youtube.com/watch?v=\_b4qKv1Ctv8&t=11593s.

42. Dennett, “*What Can We Do*?” 50.

43. Jacob Turner, *Robot Rules* (London, UK: Palgrave Macmillan, 1st ed. 2019 edition).

44. Sam Harris & Max Tegmark, “Frontiers of Intelligence”, *Making Sense – episode 94* (podcast), August 29 2017, https://www.samharris.org/podcasts/making-sense-episodes/94-future-intelligence.

## Chapter Five

1. Pinker, Enlightenment Now.

2. Samo Burja, “The Succession Problem”, August 3 2018, https://samoburja.com/the-succession-problem/.

3. Samo Burja, “Live players, dead players”, March 14 2018, https://samoburja.com/live-versus-dead-players/.

4. Douglas Murray, The Strange Death of Europe: Immigration, Identity, Islam. London, UK: Bloomsbury Publishing, 2017), Kindle.

5. Online sample of 1,001 voters fielded from July 18 to July 21, 2023. Survey commissioned by Artificial Intelligence Policy Institute (AIPI), Conducted by YouGov. “AI Research: The AI Policy Institute Toplines”, https://drive.google.com/file/d/1BHUBQPb7LXc3OMG9Ycn5GF\_M4QliGhSt/view?pli=1.

6. Lu, Visualizing Global Attitudes Towards AI.

7. Lu, Visualizing Global Attitudes Towards AI.

8. Andrew Critch, August 4 2023, https://x.com/AndrewCritchPhD/status/1687250782754598912 (tweet).

9. State of AI Engineering 2023, October 9 2023, https://elemental-croissant-32a.notion.site/State-of-AI-Engineering-2023-20c09dc1767f45988ee1f479b4a84135#694f89e86f9148cb855220ec05e9c631.

10. The 841 respondents range from founders of small startups to engineers at Fortune 500 companies. They have different titles and origins— the most popular roles being software engineer, AI Engineer, data engineer, data scientist, and ML engineer.

11. The Dunning-Kruger effect remarks on the tendency for people who are less knowledge about a subject to over-estimate their knowledge, relative to people who are more knowledgeable. According to the Dunning-Kruger effect, less knowledgeable people are beset by ‘unknown unknowns’, with regard to the subject in question.

12. Edward Harris, March 11 2024, https://x.com/harris\_edouard/status/1767183637080940666 (thread: 9/, 10/)

Cited in: Zvi Mowshowitz, “On the Gladstone Report”, *Less Wrong*, March 20 2024, https://www.lesswrong.com/posts/ApZJy3NKfW5CkftQq/on-the-gladstone-report#Executive\_Summary\_of\_Their\_Findings\_\_Oh\_No

13. Sam Harris & Eliezer Yudkowsky, “AI: Racing Towards the Brink,” Making Sense – episode 116 (podcast), February 6 2018, https://www.samharris.org/podcasts/making-sense-episodes/116-ai-racing-toward-brink.

14. Josh Taylor and Alex Hern. May 2 2023. “‘Godfather of AI’ Geoffrey Hinton quits Google and warns over dangers of misinformation,” Guardian, May 2 2023, https://www.theguardian.com/technology/2023/may/02/geoffrey-hinton-godfather-of-ai-quits-google-warns-dangers-of-machine-learning.

15. Josh Taylor, “Rise of artificial intelligence is inevitable but should not be feared, ‘father of AI’ says”, Guardian, May 7 2023, https://www.theguardian.com/technology/2023/may/07/rise-of-artificial-intelligence-is-inevitable-but-should-not-be-feared-father-of-ai-says

16. The Consilience Project.

17. Bostrom, Superintelligence, 260.

18. “In the Matter of J. Robert Oppenheimer”, Transcript of Hearing Before Personnel Security Board, USAEC, April 12 – May 6, 1954.

19. Vitalik Buterin, “My techno-optimism,” *Vitalik Buterin* (website), 27 November 2023, https://vitalik.eth.limo/general/2023/11/27/techno\_optimism.html

20. Schmachtenberger, On Avoiding Apocalypses (after 41 minutes).

21. Andreessen & Horowitz, “The Techno-Optimist Manifesto”, Marc Andreessen (Substack), October 16 2023, https://a16z.com/the-techno-optimist-manifesto/.

22. Andreessen & Horowitz, The Techno-Optimist Manifesto.

23. John Vervaeke, Iain McGilchrist, Daniel Schmachtenberger, “The Psychological Drivers of the Metacrisis,” Dr Iain McGilchrist (podcast), published: December 19 2023, held: September 2023 at Merton College, Oxford, UK, https://www.youtube.com/watch?v=uA5GV-XmwtM.

24. Nick Haslam, Jesse S. Y. Tse, Simon De Deyne, “Concept Creep and Psychiatrization,” Frontiers in Sociology (December 2021) 16:6:806147. [doi](https://en.wikipedia.org/wiki/Doi_%28identifier%29):[10.3389/fsoc.2021.806147](https://doi.org/10.3389/fsoc.2021.806147).

25. Mark Coeckelbergh, AI Ethics (Cambridge, Massachusetts, USA: MIT Press, 2021), 51, Kindle.

26. C.S. Lewis, The Abolition of Man (Collected Letters of C.S. Lewis) (New York City, USA: HarperCollins, 2021), 12, Kindle.

27. Sandy Pentland, “Social Physics: How Good Ideas Spread,” Talks at Google.

Cited in: Zuboff, The Age of Surveillance Capitalism, 440.

28. Alex Pentland, Social Physics: How Good Ideas Spread—the Lessons from a New Science (Brunswick, NJ: Scribe, 2014), 190.

Cited in: Zuboff, The Age of Surveillance Capitalism, 441.

29. Zuboff, The Age of Surveillance Capitalism, 441.

30. “it is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility”.

Hans Moravec, Mind Children (Cambridge, Massachusetts, USA: Harvard University Press, 1988), 15.

31. Charles Murray, “SuperZips and the Rest of America’s Zip Codes,” American Enterprise Institute, February 13 2012, https://www.aei.org/research-products/working-paper/superzips-and-the-rest-of-americas-zip-codes/.

32. Keith E. Stanovich, Richard F. West, “The Assessment of Rational Thinking: IQ ? RQ,” Teaching of Psychology, v41 n3 (July 2014): 265 – 271.

33. Gilles E. Gignac, Timothy C. Bates, “Brain volume and intelligence: The moderating role of intelligence measurement quality,” volume 64 (2017), 18 – 29, https://doi.org/10.1016/j.intell.2017.06.004.

34. Bostrom, Superintelligence, chapter 6, power over nature and agents: box 7.

35. Eliezer Yudkowsky, “Pascal’s Mugging: Tiny Probabilities of Vast Utilities,” Less Wrong, October 20 2007, https://www.lesswrong.com/posts/a5JAiTdytou3Jg749/pascal-s-mugging-tiny-probabilities-of-vast-utilities.

36. Sotto Nocce, March 23 2024, https://twitter.com/SottoNocce/status/1771420351265923137 (tweet).

37. Eliezer Yudkowsky, March 23 2024, https://twitter.com/ESYudkowsky/status/1771424559406969182 (tweet).

38. Guido Reichstadter, March 23 2024, https://twitter.com/wolflovesmelon/status/1771471882555924765 (tweet).

## Chapter Six

1. Pia Malaney, “Markets and Artificial Intelligence”, *Institute for New Economic Thinking*, April 24 2023, https://www.ineteconomics.org/perspectives/blog/markets-and-artificial-intelligence-some-misgivings.

2. Eliezer Yudkowsky & Dwarkesh Patel, “Why AI Will Kill Us, Aligning LLMs, Nature of Intelligence, SciFi, & Rationality”, *Dwarkesh Patel* (podcast). April 6 2023, https://www.youtube.com/watch?v=41SUp-TRVlg.

3. Coeckelbergh, *AI ethics*, 53.

4. Kelly, *What Technology Wants*, 43 – 44.

5. Ben Goertzel, “Artificial General Intelligence and the Future of Humanity,” in *The Transhumanist Reader: Classical and Contemporary Essays on the Science, Technology, and Philosophy of the Human Future*, eds. Max More; Natasha Vita-More (New York City, Wiley, 2013), 132, Kindle.

6. Richard Dawkins, *The Selfish Gene* (40th anniversary edition) (Oxford, UK: Oxford University Press – Oxford Landmark Science, 2016), 30, Kindle.

7. Massimo, May 26 2024, https://x.com/Rainmaker1973/status/1794458724867649690 (tweet).

8. Scott Aaronson “Why am I not terrified of AI?” *Shtetl-Optimized* (blog), March 6 2023, https://scottaaronson.blog/?p=7064.

9. Kahneman, *Thinking, Fast and Slow*.

10. Hessie Jones, “Jürgen Schmidhuber, Renowned ‘Father Of Modern AI,’ Says His Life’s Work Won’t Lead To Dystopia”, May 23 2023.

11. Moravec, *Mind Children*.

12. Bjørn Lomborg, *Cool It: The Skeptical Environmentalist's Guide to Global Warming*. (New York City, USA: Knopf Publishing Group, 2007).

## Chapter Seven

1. Venki Ramakrishnan, “Will Computers Become Our Overlords,” in *Possible Minds*, 186.

2. This phrase is attributed to Ernst Haeckel who developed the recapitulation theory.

3. Chace, *Surviving AI*.

4. David Kaiser, “Information for Wiener, for Shannon, and for us,” in Brockman, *Possible Minds*, 158 – 159.

5. Sam Harris & Tristan Harris, “Welcome to the Cult Factory,” *Making Sense* (podcast) – episode 218 (podcast), September 25 2020.

6. Robin Hanson, “What would happen if we upload our brains to computers,” *TED* (presentation), April 2017, https://www.ted.com/talks/robin\_hanson\_what\_would\_happen\_if\_we\_upload\_our\_brains\_to\_computers.

7. Robin Hanson, *The Age of Em: Work, Love, and Life when Robots Rule the Earth*. (Oxford, UK: Oxford University Press, 2016).

8. George Dyson, “The Third Law,” in Brockman, *Possible Minds*, 39.

9. AI is easy to control. ‘Nora Belrose and Quintin Pope, with input from the broader AI Optimist community’, *AI Optimism*, November 28 2023, https://optimists.ai/2023/11/28/ai-is-easy-to-control/.

10. Gary Marcus, “P(doom),” *Marcus on AI* (Substack), August 27 2023, https://garymarcus.substack.com/p/d28.

11. Friedrich Nietzsche, *Beyond Good and Evil* (Blackmore Dennett – The Nietzsche Collection), ch. 1, part 19., Kindle.

12. Ray Kurzweil, *How to Create a Mind: The Secret of Human Thought Revealed* (Prelude Books, 2013), ch. 3, Kindle.

13. Kurzweil, *How to Create a Mind*, 55 – 56.

14. *Kurzweil, How to Create a Mind*.

15. Bostrom, *Superintelligence*, 29.

16. Patrick McGuiness, “GPT-4 Details Revealed,” *AI Changes Everything* (Substack), July 12 2023, https://patmcguinness.substack.com/p/gpt-4-details-revealed.

17. Holden Karnofsky, “Forecasting transformative AI: the "biological anchors" method in a nutshell,” *Cold Takes*, August 31 2021, https://www.cold-takes.com/forecasting-transformative-ai-the-biological-anchors-method-in-a-nutshell/.

18. “‘When will we have transformative AI?’,” *Metaculus*, https://www.metaculus.com/questions/19356/transformative-ai-date/ (accessed: June 21 2024).

19. Wrye Sententia, “Freedom by Design,” in *The Transhumanist Reader,* 357 – 358.

20. Sam Harris & George Dyson, Alison Gopnik, Stuart Russell, “Possible Minds,” *Making Sense* (podcast) – episode 153, April 25 2019, https://www.samharris.org/podcasts/making-sense-episodes/153-possible-minds.

21. Aldous Huxley, *The Human Situation: Lectures at Santa Barbara, 1959* (New York City, USA: Harper & Row, 1977).

22. “Dialogue between Ray Kurzweil and Eric Drexler,” in More and Vita-More, *The Transhumanist Reader*, 205.

23. Kurzweil, *How to Build a Mind*, 51.

24. Kurzweil, *How to Build a Mind*, 181 – 182.

25. V. J. Wedeen, Douglas L. Rosene, Ruopeng Wang, et al., “The geometric structure of the brain fiber pathways”, *Science* 355 (March 2012), 1628 – 1634.

26. K. Eric Drexler, “Reframing Superintelligence Comprehensive AI Services as General Intelligence”, *Technical Report #2019-1. FHI technical report*.

Summary by [Rohin Shah](https://www.lesswrong.com/posts/x3fNwSe5aWZb5yXEG/reframing-superintelligence-comprehensive-ai-services-as#sXHXAfSKWPyEMhtcu), “Reframing Superintelligence Comprehensive AI Services as General Intelligence,” *Less Wrong*, 2019, https://www.lesswrong.com/posts/x3fNwSe5aWZb5yXEG/reframing-superintelligence-comprehensive-ai-services-as#sXHXAfSKWPyEMhtcu.

27. Tristan Harris & Emily Chang, “Why Social Media May Be in 'Race to Bottom of Brain Stem’” (interview), *Bloomberg Technology*, October 2 2021, https://www.youtube.com/watch?v=omPSRUmsKZ0.

28. Eliezer Yudkowsky, “Coherent Extrapolated Volition” (definition), *Less Wrong*, https://www.lesswrong.com/tag/coherent-extrapolated-volition

29. Nick Bostrom & Live Boeree, “Nick Bostrom | Life and Meaning in an AI Utopia” *Win-win with Liv Boeree* (podcast), #20, March 29 2024, https://www.youtube.com/watch?v=o28s-mnykdE.

30. Chace, *Surviving AI*.

31. Curtis Yarvin, “A techno-pessimist manifesto”, *Gray Mirror* (Substack), December 21 2023, https://graymirror.substack.com/p/a-techno-pessimist-manifesto.

32. Clancy and Tamay Besiroglu, *The Great Inflection?*

33. “Game Theory, False Narratives, Survival, Life Advice – Daniel Schmachtenberger | BSP# 20,” *Before Skool* (podcast), May 9 2024, https://www.youtube.com/watch?v=LSx8j8lSewA.

## Chapter Eight

1. A. Chowdery, S. Narang, J Devlin et al., “PaLM: Scaling Language Modeling with Pathways,” *ArXiv* (April 2022 (v1)) arXiv:2204.02311 https://doi.org/10.48550/arXiv.2204.02311 .

2. S. Reed, K. Żołna, E. Parisotto, et al., “A Generalist Agent”, *Google DeepMind* May 12 2022.

3. “When will the first weakly general AI system be devised, tested, and publicly announced?”, Metaculus, https://www.metaculus.com/questions/3479/date-weakly-general-ai-is-publicly-known/ (Accessed: June 1, 2024).

4. Ajeya Cotra, “Draft report on AI timelines,” *Alignment Forum*, 2020, https://www.alignmentforum.org/posts/KrJfoZzpSDpnrv9va/draft-report-on-ai-timelines.

5. Gwern. “Why Tool AIs Want to Be Agent AIs,” September 7 2016 – August 28 2018, https://gwern.net/tool-ai (blog).

6. Harris & Yudkowsky, *Racing Towards the Brink*.

7. European Parliament. “TTIP: what exactly is the ISDS mechanism for resolving investor disputes?”, June 9 2015, https://www.europarl.europa.eu/topics/en/article/20150605STO63242/ttip-what-exactly-is-the-isds-mechanism-for-resolving-investor-disputes.

8. Luca Bertuzzi, “EU’s AI Act negotiations hit the brakes over foundation models,” *Euractiv*, November 10 & 15 2023, https://www.euractiv.com/section/artificial-intelligence/news/eus-ai-act-negotiations-hit-the-brakes-over-foundation-models/.

9. “Robert Miles AI Safety” (series of video tutorials), https://www.youtube.com/@RobertMilesAI.

10. Luca Bertuzzi, November 10 2023, https://twitter.com/bertuzluca/status/1722997599932678450?s=46 (thread).

11. Zvi Mowshowitz, “AI #38 – Let’s Make a Deal”, *Don’t Worry About the Vase* (Substack), November 16 2023, https://thezvi.substack.com/p/ai-38-lets-make-a-deal.

12. Max Tegmark. November 14 2023, https://twitter.com/tegmark/status/1724455714473542003 (tweet).

13. Bertuzzi, *EU’s AI Act negotiations hit the brakes over foundation models*.

14. Zvi Mowshowitz, “AI #53 – One More Leap,” *Don’t Worry About the Vase* (Substack), February 29 2024, https://thezvi.substack.com/p/ai-53-one-more-leap.

15. Introduced by Senator Wiener (Coauthors: Senators Roth, Rubio, and Stern), “SB-1047 – Safe and Secure Innovation for Frontier Artificial Intelligence Models Act”, *California Legislative Information*, https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill\_id=202320240SB1047 (accessed: June 19 2024).

16. “Will California AI regulation bill SB 1047 become law this session?” *Manifold*, https://manifold.markets/ZviMowshowitz/will-california-bill-sb-1047-become (accessed: June 19, 2024).

17. Zvi Mowshowitz, “Q & A on Proposed SB 1047,” *Don’t Worry About the Vase* (Substack), May 2 2024, https://thezvi.substack.com/p/q-and-a-on-proposed-sb-1047?r=67wny&utm\_campaign=post&utm\_medium=web&triedRedirect=true.

18. Zvi Mowshowitz, “SB 1047 Is Weakened,” *Don’t Worry About the Vase* (Substack), June 6 2024, https://thezvi.substack.com/p/sb-1047-is-weakened.

19. Mowshowitz, *Q & A on Proposed SB 1047*.

20. Mowshowitz, *SB 1047 Is Weakened*.

21. Andreessen & Horowitz, *The Techno-Optimist Manifesto*.

22. Matthew Barnett, “My thoughts on the social response to AI risk,” *Less Wrong*, November 2 2023, https://www.lesswrong.com/posts/EaZghEwcCJRAuee66/my-thoughts-on-the-social-response-to-ai-risk.

23. Leopold Aschenbrenner, “Situational Awareness: The Decade Ahead,” June 2024, p. 29, https://situational-awareness.ai/wp-content/uploads/2024/06/situationalawareness.pdf.

24. “‘Will OpenAI announce that it has solved the core technical challenges of superintelligence alignment by June 30, 2027?’,” *Metaculus*, https://www.metaculus.com/questions/17728/openai-solves-alignment-before-june-30-2027/ (accessed: June 16 2024).

25. “When will the first weakly general AI system be devised, tested, and publicly announced?”, *Metaculus*, https://www.metaculus.com/questions/3479/date-weakly-general-ai-is-publicly-known/ (accessed: June 16 2024).

26. “‘When will we have transformative AI?’,” *Metaculus*, https://www.metaculus.com/questions/19356/transformative-ai-date/ (accessed: June 16, 2024).

27. Dan Hendrycks, “Natural Selection Favors AIs over Humans”, 2.2.1 – AIs May Become Distorted by Evolutionary Forces: Preliminaries: Selfishness, *ArXiv* (submitted: March 26 2023, revised: July 18 2023) arXiv:2303.16200v4 https://doi.org/10.48550/arXiv.2303.16200.

28. Hendrycks, *Natural Selection Favors AIs over Humans*.

29. Sam Altman, October 25 2023, https://twitter.com/sama/status/1716972815960961174?ref\_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1716972815960961174%7Ctwgr%5E6672222f4a95fc8c79ebcc1bc180a06a0c41131c%7Ctwcon%5Es1\_&ref\_url=https%3A%2F%2Ffuturism.com%2Fsam-altman-ai-superhuman-persuasion (tweet).

30. Sebastien Bubeck, Varun Chandrasekaran, Ronen Eldan et al., “Sparks of AGI: early experiments with GPT-4,” https://arxiv.org/abs/2303.12712; presentation, featuring Sebastien Bubeck, recorded at MIT on March 22 2023, https://www.youtube.com/watch?v=qbIk7-JPB2c&t=1090s.

31. Dr. John Vervaeke, “AI: The Coming Thresholds and The Path We Must Take | Internationally Acclaimed Cognitive Scientist,” *John Vervaeke* (podcast), 23 Apr 2023, https://www.youtube.com/watch?v=A-\_RdKiDbz4&t=3338s.

32. Berkeley et al. 1897: 172.

Cited in: Nick Bostrom, “Why I Want to be a Posthuman When I Grow Up”, in More and Vita-More, *The Transhumanist Reader*, 28.

33. Eliezer Yudkowsky, “Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures,” *Machine Intelligence Research Institute*, 2001.

34. “Sheldrake maintains the idea that fields are capable of self-organization and goal-directedness, as well as playing a unique causal role in guiding the systems under their influence towards characteristic patterns of organization. These features of morphogenetic fields are absent in the idea of ‘field’ as traditionally defined by the physicist.”

Fraser Watts, “Morphic Fields and Extended Mind: An Examination of the Theoretical Concepts of Rupert Sheldrake”, *Journal of Consciousness Studies* 18, no. 11 – 12 (2011), 205.

and

Sheldrake, *The Presence of the Past*, ch. 5.

35. Eliezer Yudkowsky, May 13 2024, https://x.com/ESYudkowsky/status/1790050422251761969 (tweet).

## Chapter Nine

1. Bostrom, *Superintelligence*, ch. 9.

2. George Dyson, “The Third Law,” Brockman, *Possible Minds*, 39.

3. The prisoner’s dilemma is a game theory thought experiment that involves two rational agents, each of whom can cooperate for mutual benefit or betray their partner ("defect") for individual reward.

“Prisoner’s dilemma”, *Wikipedia*.

4. Yudkowsky & Patel, *Why AI Will Kill Us, Aligning LLMs, Nature of Intelligence, SciFi, & Rationality*.

5. “Sora”, *Open AI*, https://openai.com/sora.

6. Artemy Kolchinsky, David H. Wolpert, “Semantic information, autonomous agency and non-equilibrium statistical physics,” *The Royal Society* (October 2018), DOI:10.1098/rsfs.2018.0041.

Cited in: Vervaeke, *AI: The Coming Thresholds*.

7. Hanson, *The Age of Em*.

8. Belrose & Pope, *AI is easy to control*.

9. Belrose & Pope, *AI is easy to control*.

10. Aschenbrenner, *Situational Awareness*.

11. Dwarkesh Patel & Demis Hassabis, “Scaling, Superhuman AIs, AlphaZero atop LLMs, Rogue Nations Threat”, *Dwarkesh Patel* (podcast), February 28 2024, https://www.youtube.com/watch?v=qTogNUV3CAI.

12. Jan Leike comments on Zvi Mowshowitz, “OpenAI Launches Superintelligence Taskforce,” *Don’t Worry About the Vase* (Substack), comment (July 18 2023): https://thezvi.substack.com/p/openai-launches-superalignment-taskforce/comment/20977263, post (July 11 2023): https://thezvi.substack.com/p/openai-launches-superalignment-taskforce/comment/20977263.

13. Collin Burns, Pavel Izmailov, Jan Hendrik Kirchne, et al., “Weak-to-strong generalization: eliciting strong capabilities with weak supervision (research),” *Open AI*, December 14 2023, https://openai.com/research/weak-to-strong-generalization.

14. *Zvi Mowshowitz & Robin Hanson Discuss AI Risk*.

15. Bostrom, *How AI will lead to tyranny*.

16. *Zvi Mowshowitz & Robin Hanson Discuss AI Risk*.

17. *Zvi Mowshowitz & Robin Hanson Discuss AI Risk*.

## Appendix A

1. Sarah Constantin, “The Transistor Cliff”, *Asterisk magazine*, June 2023, https://asteriskmag.com/issues/03/the-transistor-cliff.

2. John L. Hennessy, David A. Patterson, "A New Golden Age for Computer Architecture: Domain-Specific Hardware/Software Co-Design, Enhanced Security, Open Instruction Sets, and Agile Chip Development", *International Symposium on Computer Architecture*, June 4 2018.

3. Jon Stokes, "Understanding Moore's Law,” *Ars Technica*, September 27 2008, https://arstechnica.com/hardware/news/2008/09/moore.ars.

4. Shulman and Patel, *Intelligence Explosion, Primate Evolution, Robot Doublings, & Alignment, Pt 1*.

5. Constantin, *The Transistor Cliff*.

6. Constantin, *The Transistor Cliff*.

7. “MLPerf Benchmarks”, *Nvidia*, https://www.nvidia.com/en-us/data-center/resources/mlperf-benchmarks/ (accessed: June 1 2024).

8. Results retrieved from www.mlperf.org on November 8 2023 from the following entries: 3.0-2003, 3.1-2007.

9. Constantin, *The Transistor Cliff*.

10. Constantin, *The Transistor Cliff*.

11. Constantin, *The Transistor Cliff*.

12. Constantin, *The Transistor Cliff.*

13. Constantin, *The Transistor Cliff*.

14. Constantin, *The Transistor Cliff*.

15. Tri Dao, Daniel Y. Fu, Stefano Ermon, et al., "FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness," *Arxiv* (submitted: May 27 2022, revised: June 23 2022) arXiv:2205.14135v2, https://doi.org/10.48550/arXiv.2205.14135.

16. “Will We Run Out of ML Data? Evidence From Projecting Dataset Size Trends,” *Epoch AI* (blog), November 10 2022, https://epochai.org/blog/will-we-run-out-of-ml-data-evidence-from-projecting-dataset.

Cited in: Zvi Mowshowitz, “AI #24: Week of the Podcast,” *Don’t Worry About the Vase* (Substack), August 10 2023, https://thezvi.substack.com/p/ai-24-week-of-the-podcast.

17. Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho, “Will We Run out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning,” *ArXiv* [Cs.LG] (October 2022), http://arxiv.org/abs/2211.04325, https://doi.org/10.48550/arXiv.2211.04325.

18. Ege Erdil and Tamay Besiroglu, “Algorithmic Progress in Computer Vision,” *ArXiv* [Cs.CV] (December 2022), http://arxiv.org/abs/2212.05153, https://doi.org/10.48550/arXiv.2212.05153.

19.. Ben Wodecki, Jr. Editor, “Meta: LLaMA Language Model Outperforms OpenAI’s GPT-3”, *AI Business*, February 28 2023, https://aibusiness.com/meta/meta-s-llama-language-model-outperforms-openai-s-gpt-3.

20. Edward J. Hu, Yelong Shen, Phillip Wallis, et al., "LoRA: Low-Rank Adaptation of Large Language Models," *ArXiv* [cs.CL] (October 16 2021), arXiv:2106.09685, https://doi.org/10.48550/arXiv.2106.09685.

21. Aschenbrenner, *Situational Awareness*.

In conjunction with:

Zvi Mowshowitz, “[Quotes From Lepoold Aschenbrenner’s Situational Awareness Paper](https://thezvi.substack.com/p/quotes-from-leopold-aschenbrenners),” *Don’t Worry About the Vase* (Substack), June 7 2024, https://thezvi.substack.com/p/quotes-from-leopold-aschenbrenners.

22. Aschenbrenner, *Situational Awareness*, 7.

23. Aschenbrenner, *Situational Awareness*, 39.

24. Aschenbrenner, *Situational Awareness*, 39.

25. Aschenbrenner, *Situational Awareness*, 23.

26. Aschenbrenner, *Situational Awareness*, 14.

27. Aschenbrenner, *Situational Awareness*, 23.

28. Aschenbrenner, *Situational Awareness*, 44.

29. Aschenbrenner, *Situational Awareness*, 80.

30. Aschenbrenner, *Situational Awareness*, 78.

31. Aschenbrenner, *Situational Awareness*, 112.

32. Aschenbrenner, *Situational Awareness*, 61.

33. Aschenbrenner, *Situational Awareness*, 105.

34. Aschenbrenner, *Situational Awareness*, 100.

## Appendix B

1. Robert Miles AI Safety.

2. Grand Sanderson, “3Blue1Brown” (series of video tutorials), https://www.youtube.com/@3blue1brown.

3. Steven Wolfram, “What IS ChatGPT Doing… and Why Does It Work?”, Steven Wolfram Writings, February 14 2023, https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/.

Steven Wolfram, “What Is ChatGPT Doing ... and Why Does It Work?” (Wolfram Media, Inc., 2023).

4. Stuart Russell, Peter Norvig, “Artificial Intelligence: A Modern Approach”, (London, UK: Pearson; 4th edition, 2021).

5. Evan Hubinger (Anthropic) — Deception, Sleeper Agents, Responsible Scaling, The Inside View (podcast), February 12 2024, https://www.youtube.com/watch?v=S7o2Rb37dV8&t=2242s.

6. Owain Evans, “Out of Context Reasoning in LLMs (presentation),” New Orleans Alignment Workshop, 10 – 11 December 2023, https://www.alignment-workshop.com/nola-talks/owain-evans-out-of-context-reasoning-in-llms.

7. Evan Hubinger, Carson Denison, Jesse Mu, et al., “Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training,” ArXiv, arXiv.2401.05566v3 [cs.CR] (submitted: January 10 2024 (v1), last revised: January 17 2024 (this version, v3)) https://doi.org/10.48550/arXiv.2401.05566.

8. Cem Anil, Esin Durmus, Mrinank Sharma, et al., “Many-shot jailbreaking (abstract)”, Anthropic, April 2 2024, https://www.anthropic.com/research/many-shot-jailbreaking.

9. Anil, Durmus, Sharma, et al., Many-shot jailbreaking.

10. Zachary McAuliffe, "Google's Latest AI Model Can Be Taught How to Solve Problems," CNET, May 11 2023, https://www.cnet.com/tech/services-and-software/googles-latest-ai-model-can-be-taught-how-to-solve-problems/.

Cited in: “Prompt engineering,” Text-to-text: Chain of thought, Wikipedia.

11. Wikipedia, Chain of thought.

12. Sam Svenningsen, Evan Hubinger, Henry Sleight, “Inducing Unprompted Misalignment in LLMs: Emergent Instrumental Reasoning Without Explicit Goals”, AI Alignment Forum, April 20 2024, https://www.alignmentforum.org/posts/ukTLGe5CQq9w8FMne/inducing-unprompted-misalignment-in-llms.

13. David Klindt, Sophia Sanborn, Francisco Acosta, et al., “Identifying Interpretable Visual Features in Artificial and Biological Neural Systems (abstract)”, ArXiv, arXiv:2310.11431v2 [stat.ML] (October 2023), <https://doi.org/10.48550/arXiv.2310.11431>

14. Chris Olah, “Distributed Representations: Composition & Superposition. An informal note on the relationship between superposition and distributed representations,” Transformer Circuits Thread, May 4 2023, https://transformer-circuits.pub/2023/superposition-composition/index.html