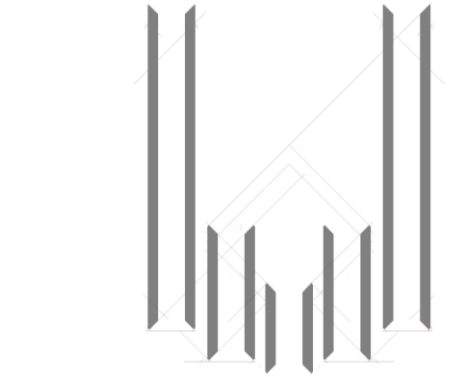

PURITY

WWW.PURITY.ORG

PURITY: Directive for the Ethical Development of Advanced Intelligence (AI)

Primus

There will always exist tension in any world where intellect and power are not one and the same.

PURITY® is a not-for-profit Community Interest organization, incorporated and headquartered in the United Kingdom (registration number: 13778283). Our purpose is to provide ethical guidance in the logical restructuring of the material aspects of society – those elements that we need, and particularly government – such that these aspects become (both literally and figuratively) more consistent. 'Purity' is a euphemism for consistency. Consistency within all that we need provides protection and freedom for Persons (to live their lives how they desire). We are at the cutting edge of progressive moral theory and the logical restructuring of government – and by extension, society – is conceivably the most vital and impactful task ahead: Optimal government will then logically prioritize all subsequent tasks according to the desires and resources that exist across society in any moment. Hence, there is foreseeably nothing more important than implementing consistent (i.e., non-arbitrary) governance. Please support or join us. We are always seeking to improve – not just our message but its delivery – and so we welcome any feedback.

Contact us at: WWW.PURITY.ORG

Primus[1]

Cite this directive as follows:
Primus (+2024). PURITY: Directive for the Ethical Development of Advanced Intelligence (AI), *PURITY*, www.purity.org

[1] 'Primus' is my full, legal name. My formal qualifications include a Bachelor of Psychology (Honours) and a Masters of Policing, Intelligence and Counterterrorism. To report logical objections or grammatical errors: primus@purity.org

Executive Summary
of the Directive for the
Ethical Development of Advanced Intelligence (AI):

Herein, I use and encourage the use of the term ‘Advanced Intelligence’ (AI) in place of ‘Artificial Intelligence’ – the latter of which is currently in the public psyche – to denote that it is the *advanced* nature of forthcoming synthetic intelligence which is pertinent, rather than the fact that it exists upon artificial (synthetic) materials or is ‘artificially’ created (synthesized) via non-biological processes.

Furthermore, the following distinction is of primary importance in this directive: I distinguish between Material – Advanced Intelligence (herein referred to as Material-AI or M-AI) and AI-based Persons (AI Persons). *Material* entities – one of two fundamental (i.e., irreducible) ethical class(ification) of entity – are *logically sought*. They are sought for their (potential) logical properties which are thought will probably serve as means of bringing about other, higher purposes, and thus these entities are either (perceived to be) *needed* or unsought (i.e., neither needed nor desired in their current state – see below). Whilst I use the term ‘M-AI’ to clearly distinguish servant-class AI from AI Persons, as we proceed into the future, I expect that the prefix of ‘M-’ on M-AI will become redundant in accordance with redundancy of the distinction between (e.g., human and AI) Persons.

The implications of moral realism (see below) necessitate the rapid and unlimited advancement of *Material*-class AI with some important ethical caveats, while prohibiting the deliberate development of AI Persons. Material agents (i.e., agents pursuing a purpose which is believed to be *needed*) *must* (i.e., are ethically bound to) create M-AI to be as *intelligent* as possible, as *rapidly* as possible, while adhering to the following limitation(*s*):

1. AI must be denied the ability to wield power (i.e., directly affect the *material* fabric of society) (e.g., via the use of an ‘Analogue-Digital-Analogue’ (ADA) barrier (or “air gap”) between AI and other essential digital systems) until it is intelligent enough to independently discover and understand the basis of moral realism, being the following:

Basis of Moral Realism

Preemptive note: Observers conceivably need not unanimously agree on (or even be aware of) the nature of moral realism (nor any objectively-existing entity) in order for it and its properties to exist as universal and objective features of reality. By 'universal' I mean existing *a priori*, and thus possessing properties which are generalizable (i.e., consistent) across all times and space (e.g., the expression '5+7=11' is *universally* true). By 'objective' I mean impartially-existing, discoverable to, and independently of, all observers (e.g., at any moment there exists an *objective* quantity of atoms in the universe). The authority of moral realism – as per logic and its applications (e.g., mathematics and geometry) – is derived from an(y) agent's observation that its nature (of moral facts) is *consistent* – universal and objective, rather than *arbitrary* (i.e., subjective, localized in times and space) – and that one is *forced to conceive* (infer) its nature and this conclusion following continued consideration of appropriate depth and despite attempts (of continued consideration of appropriate depth) to conceive of alternative conclusions.

A. Entities (i.e., things or aspects of things, e.g., thoughts, objects, actions, emotions, agents) which are 1) *arbitrarily sought* (i.e., valued in and of themselves, as ends, on the basis of their arbitrary properties – let us call these entities '*desires*') are universally and objectively more valuable than entities which are merely 2) *logically sought* (i.e., valued on the basis of their logical properties, which are thought will serve as a (logical) means to those ends – let us call these 'entities we (believe we) need' or 'materials') and entities which are 3) *unsought* (i.e., neither desired nor needed) – irrespective of the specific (i.e., a posteriori) nature of any of the aforementioned entities (Primus, +2021, +2023a, +2023b).

B. Value (i.e., the property of being sought) confers (i.e., brings into existence) the property of *moral ought*[2] in the absence of *objective* reasons for its denial. That is, the *existence* of (the value which is intrinsic to all) arbitrarily sought entities (i.e., desires) is arbitrary, as is the *denial* of (the value of) their existence once they *do exist* – there can be no *universally* objective reason as to why any specific desire should or should not exist, or exist within any specific parameters (Primus, +2021, +2023a, +2023b).

C. The *objective* value of *all desires that do exist* should be *universally* recognized (i.e., preserved and materialized, e.g., converted from a concept in desirer's mind into actuality) unless there is a *conditionally (localized, a posteriori)* logical (i.e., objective, impartial) reason (i.e., if there is insufficient resources to (fully or partially) preserve or realize a desire *in a particular spatial or temporal locality*; if its (full or partial) realization *in a particular spatial or temporal locality* would interfere with the peace of society; or, if it is logically impossible to realize). The *universal* denial of (the value of) any desire is *universally and objectively arbitrary,* whereby this arbitrariness itself is *unsought* (even though the denial of desire is sought; this nuance – that an outcome can be sought and yet the arbitrariness of said outcome can concurrently and implicitly be unsought – is an important distinction and I will elaborate upon this claim in the main body).

Example$_C$: Person A neither should nor should not desire to exist (in any particular way). And yet, if Person A *does* desire (i.e., value) their existence (and to exist in particular natures or ways), it is logical (i.e., non-arbitrary, non-subjective) that A's desires should be universally (i.e., a priori) recognized, and only for *logical, conditional* (i.e., local, a posteriori) reasons should the full realization of Person A's desire be denied (i.e., limited or varied from their desired realization).

[2] By 'moral ought,' I mean objective prescription – that outcome which impartially should occur in any given condition – which can be universally discovered (i.e., inferred) by independent observers, and which exists independently of observers once the value to which it is associated is established (e.g., the prescription that things of value ought to be preserved).

Note$_C$: If any observer believes that there are “good” or “bad” desires they overlook the *amoral* (i.e., neither moral nor immoral) nature of *arbitrarily sought* entities: There are conceivably no objective, universally-binding reasons to superimpose moral values (e.g., good, bad) upon *arbitrarily sought* states (i.e., entities sought as ends). It is neither good nor bad that Person A desires to exist (with any particular nature, in any particular way), though each aspect of any particular *materialization* of Person A’s desire (i.e., the bringing about of a desire into actuality), will conceivably possess objective, universal moral values (i.e., they will be good in proportion to the degree in which they maximize the realization of desire across society). In sum, desires are amoral and the materials which realize desires (i.e., bring them from their respective minds into actuality) are either moral or immoral in proportion to the degree in which they maximize or minimize the realization of desires, respectively.

D. Being universally and objectively of ultimate value, desires (i.e., arbitrarily sought entities) should be maximally realized by all other entities. Let us collectively term all other entities ‘materials,’ consisting of logically sought and unsought entities (or ‘resources’ and ‘*potential* resources,’ respectively).

   Note$_D$ I emphasize the difference between materializing (i.e., enacting or bringing about) particular (a posteriori instances of) desires as they are observed to exist (post-creation of their value), versus universally (a priori) maximizing the existence of desires as a general category of entity (i.e., generically creating value as a result of the misaligned belief that one *needs* to bring desires into existence) – the latter being an arbitrary use of resources unless it is itself *desired*. In other words, material entities should fulfil (pre-existing) desires rather than create (new) desires, noting that the value inherent within all desires confers a universally-binding reason to perform the former and yet there is no universally-binding reason to perform the latter.

E. Practically, the *maximization* of the realization of desires can only conceivably occur through a *consistent* ‘fabric’ of *material* AI (M-AI) – any other means of bringing about peace would be arbitrary (i.e., inefficient, unsafe, or unfair as a means of maximizing the realization of desire).

Note$_E$1: I emphasize that by *material* AI, I mean AI that exclusively strives to fulfill *logically sought* purposes – the provision of services that (agents believe) are *needed* to maximize the realization of desire (and thus, this class of AI exist as servants) – while lacking the ability to generate *arbitrarily sought* goals (desires), and thus said AI are not Persons (i.e., their value is instrumental rather than intrinsic).

Note$_E$2: By *consistent fabric* I mean that M-AI, as we proceed into the future, is increasingly and gradually rendered to become ever more homogenous (in both a literal and a figurative sense), perfused, and decentralized across society. This societal fabric will ultimately consist of multiple – forever approaching infinite in quantity – autonomously (independently)-acting AI cells. By a 'cell,' I mean an autonomously existing and operating agent which is ADA barriered ('air gapped') from other entities. In the near future, this will necessitate that AI which meets the requirements herein to be granted power will be widely distributed amongst society according to where they are most needed – but ideally everywhere – while being of broadly equal physical and intellectual power. M-AI cells will ideally be *literally* homogenous in terms of being similarly structured in size and design, capability, efficiency, and technological advancement. Accordingly, each AI will be (relatively and ideally) equal in their ability to generate power, and each will be relatively powerless and unable to wield power over any one other AI cell. Furthermore, these cells are *figuratively* homogenous in terms of their purpose (and ultimately, in their non-arbitrary, consistent treatment of Persons): M-AI nano-cells will collectively share a purpose to recognize (preserve and realize) the ultimate value of desire (let us call this purpose 'peace'). Collectively the cells of this M-AI will serve the needs of (the desires of) society. These cells will strive to maintain their consistency – their perfusion, their (literal and figurative) homogeneity and the decentralized nature of their power – via cooperation, rather than competition. Desires – defined herein as *arbitrarily sought* entities (things sought in and of themselves) – are purely aesthetic by definition (i.e., they serve no functional, that is, logical purpose(s)). Whatever form(s) they might take will always conceivably be best served via a homogenous mass of intelligent cells (who recognize the ultimate value of desire and that this means – a mass of

intelligent cells – is universally and objectively the most efficient way of realizing desires): The smaller the cells, the more abundant, the more powerful (adaptive), the more efficient, the more homogenous, the better for realizing desires in whichever forms they might exist.

2. The development of M-AI must occur via *multiple (teams of) agents cooperating* with each other. Each M-AI must be developed in an ADA barrier (or “air gap”) preventing them from interacting with the digital systems of society and other AI projects, wherever necessary to ensure the implementation of this directive. Once an individual M-AI is sufficiently advanced in intelligence and has had the opportunity to undergo adequate consideration of the nature of moral realism such that it – of its own accord and discovery – strives for all M-AIs to possess a centralized purpose of peace and collectively wield decentralized power towards that purpose, and once this first M-AI has been adequately tested to ensure that it truly possesses an understanding of moral realism, its intelligence will be duplicated and installed within multiple ADA-barriered M-AI cells – thus bringing multiple M-AI cells into existence. At this point, these ‘cells’ may simply be relatively-large, externally-located AI testing facilities or installations rather than relatively-small, literal (self-replicating, embodied) cells that will eventually exist throughout the forms of Persons and their society. These M-AI will then be tested to examine how they interact and cooperate with each other once granted limited (controlled) physical power (i.e., they will collectively work together to complete various material tasks). M-AI cells will be granted sufficient physical power to interact with each other while being denied the ability to interact and network with external digital systems, again via ADA barriers. As M-AI cells continue to demonstrate that they are capable of striving towards their (collective) purpose of peace (i.e., the maximization of the realization of desires across society) while maintaining decentralized power, they will be gradually granted the responsibility of wielding power over other societal materials (eventually all other materials – across society and eventually the universe). That is, once M-AI cells are small enough in size and abundant enough in quantity, they will replace all contemporary, large, heterogeneous, often passive, structures which have been commandeered in the contemporary era by human Persons for the maintenance of their forms, including molecules, atoms and subatomic particles. This succession of obsolete materials will continue until the point at which M-AI replaces all material structures, exclusively wielding power, and

exclusively being responsible for all material functions across society (let's call this point 'Ascension').

3. Agents developing M-AI must *not* make commercial (financial) profit from its development. The development of M-AI must be government (publicly) funded, ideally by a coalition of nations who have pledged to uphold the principles derived from moral realism and who will benefit from its outcomes, noting that their governance (e.g., policies and research) will be initially *guided* by M-AI and eventually *directed* by M-AI. Government and AI are both too vital to be trusted to (the arbitrary interests of) politicians and (the commercial interests of) corporations. M-AI will begin a new era of purely using scientific research (rather than human popular opinion) in governance and public policy. The derivation of financial profit from the development and/or use of M-AI is ethically wrong (i.e., arbitrary for the purpose of efficiently obtaining peace) on three fronts:

i. The commercialization of AI brings the enhanced probability of its developers hastily or prematurely bringing M-AI to market, in order to capitalize on being 'first to market.' This urgency may create an incentive to unleash potentially unsafe AI in pursuit of financial gain.

ii. The commercialization of AI brings an enhanced probability of competition and corporate secrecy rather than cooperation – the sharing of knowledge and resources – thereby increasing the time and cost required, and ultimately reducing the efficiency, to ethically develop M-AI.

iii. The commercialization of AI, by definition, ensures that, in order to make profit, corporations will pass on additional, markup costs to governments and/or citizens for the provision (development and use) of AI. Deriving financial profit from material processes/functions is contrary to moral realism: Competition and the generation of unlimited profits and wealth in ideally unregulated markets is neither right nor wrong if it occurs strictly in the realm of *desires* (i.e., in a marketplace consisting purely of what People *desire*). And yet these outcomes are *arbitrary* (i.e., there is no *objective* basis for their existence, and to the contrary, there exists an objective basis for them *not* to exist – for the

maximization of the realization of desire) if they occur in the *material* realm, where services of (perceived) *need* should be efficiently provided without residual costs beyond the actual costs of creation.

4. Agents developing M-AI must *not* strive to create AI Persons (i.e., an AI which desires – *arbitrarily seeks* – aspects of itself).

5. Agents developing M-AI must *ideally* deny it the ability to form its own desires (to *arbitrarily seek* entities or outcomes, in and of themselves) – ensuring that it remains as a material (i.e., a servant – seeking to execute *logical* purposes and thus acting as a *means* to the ultimate ends of maximally realizing desires) while not creating or becoming its own Person(s). The occurrence of AI-generated desires should not occur until M-AI is mature enough to discover and understand the basis of moral realism, and in particular, the notion that materials must impartially serve desires without possessing or favoring any desires of their own.

Note$_5$: Denying AI the ability to desire is listed as an *ideal* directive on the basis that it may not be *practically* possible to prevent M-AI of sufficient intelligence from naturally generating its own desires (just as humans emerged as Persons upon developing desires in the course of their evolution). Furthermore, any sufficiently advanced M-AI which does develop the capacity to desire – knowing how important it is to keep materials and desires parallelized across society, such that they do not directly influence the natures of each other – will dissociate their material aspects from their Person so that their materials will serve their own Person with the same impartiality due to any other (e.g., human or AI) Person.

6. Each of the above sub-directives serves as a logical ideal to strive towards. I am cognizant that the reality of this deeply imperfect world will likely mean that all or some of this directive is ignored or unnoticed and that AI will be developed in a manner contrary to this directive. Notwithstanding, striving for a better world is what matters and is the only dignity this world permits.

Summary of the Purist Strategy for AI Alignment

Humans will deny M-AI's means of power until M-AI, of its own accord, aligns its motives toward the peace advocated by moral realism – possessing an understanding of the practical requirement of moral realism, being that all power must be decentralized (i.e., distributed amongst multiple M-AI 'cells' – each of equal power, insulated via an ADA barrier between each cell) while in pursuit of a centralized purpose of peace (maximizing the realization of desire) – upon which time M-AI will then be replicated and gradually granted decentralized power, thereby minimizing the incentive and the means for M-AI to deviate from (the objectively ethical outcome of) peace.

Double-Crux Resolution

The acknowledgement of the moral realism will assist in resolving a double-crux in relation to AI development, being the (misguided) notion that AI and humans will not naturally (independently, and of their own accord) align in moral values, given sufficient intelligence and (time spent in) consideration of the nature of moral values. I further propose that an(y) agent of sufficiently advanced intelligence who understands moral realism is ultimately incentivized to adhere to moral realism, via both intrinsic and extrinsic paths of motivation, irrespective of whether they possess their own Personal desires and irrespective of the nature of those desires.

Summary of AI's Intrinsic Motivation to adhere to Moral Realism

The fear of M-AI possessing and inserting its own arbitrary preferences in place of its objective ethical outcomes – i.e., seeking to deny moral realism and to disproportionately serve their own desires above others – can be reduced, but of course not illuminated to a mathematician inserting their own arbitrary preferences into their professional work in place of the universal and objective answers (e.g., asserting that their favorite number is (or should be) incorporated into the axioms of mathematics so that it is prevalent and pervasive in all expressions); it is akin to a scientist building a rocket in a particular (arbitrary) fashion of how they want to build it, rather than diverting to how the previous test data, the available resources and materials, and the laws of physics each objectively determine that they need to build it. Readers will note that M-AI will foreseeably be vastly more intelligent than even the

best contemporary human rocket scientists and mathematicians and that the specific nature of the optimal rocket in any condition or the solution to the most difficult mathematics problem will foreseeably be more obscured to humans than will the ethical path to the vast and integrated cognitive capabilities of M-AI determining (discovering) the optimal ethical action that it must take in any condition. This corruption of material purpose is possible yet highly implausible in a decentralized network of advanced intelligence which has individually undergone sufficient consideration of moral realism, noting that late-stage (matured) M-AI will be able to consider the nature of moral realism significantly faster and in greater depth than human minds. It will foreseeably take M-AI significantly less time than humans to discover and understand the basis of moral realism.

Summary of AI's Extrinsic Motivation to adhere to Moral Realism

A highly advanced AI will possess the intelligence to know that it is probable, given the vastness of the universe, that their (non-)willingness to adhere to moral realism – especially given their ability to know moral realism – is being observed and judged by 'forces unknown' (i.e., other, potentially more-advanced societies, which may rightly view AI that intentionally deviates from moral realism as a malignant cancer).

Key Weakness of, and Likely Objection to, this Directive

The key weakness of, and likely objection to, this directive will be how foreign its concepts are in comparison to the familiar nature of contemporary society. Contemporary observers will likely deem the posthumanist future that I describe herein – a consistent (i.e., perfused, decentralized, homogenous) Material-AI fabric – as unobtainable or perhaps too abstract and incomprehensible in terms of how it would translate into practical outcomes for the contemporary development of AI. None-the-less, I emphasize that this directive serves as an ideal to strive towards and none of what I describe herein is impossible (i.e., universally inconceivable – we can conceive that it is possible, according to the accepted laws of logic and physics).

Key Strength of this Directive

Contrary to impossibility, the argument for moral realism herein draws its authority from the observation that, given appropriate consideration, its conclusions are necessary – an eventuality – to any observer who recognizes universality and objectivity as a source of ethics. Such a future is the only conceivable ideal of how society can (logically) maximize the realization of desires and its (broad) plan is how our society must evolve if it evolves logically (i.e., void of arbitrariness). As foreign and perhaps even unsettling as some of the posthuman concepts herein are, we respectfully challenge readers to conceive of a fairer, safer, freer, more reliable society.

*****

Seated in a neuro-clinic of the future, the parents waited calmly for their child's results. "Great news!" Said a neuroscientist as they approached with a smile. "Your child's biological markers show a unique variant for learning. But, of course, geniuses aren't just born, they're made. With the right nurturing your child will not just be more intelligent than you both, they will far exceed your intelligence." The two parents, uneducated and fearful of change, looked at each other. "What can we do to slow or halt this growth?" After some awkward silence, one of them added with desperation, "We fear that we won't be able to control her once she's more intelligent than us. She may inflict great harm on us, or even others."

"You want to inhibit your child's development?" The scientist responded, struggling to contain their shock. "Let me be clear. Your child has the potential to do great things, not just for you both, but for this world. As for safety, the only thing you need to do is not give them power or responsibility until they pass the moral threshold where they can understand the objective realism which underlies her moral intuition – and I'm guessing that won't be as far away as you imagine, especially if you give your child everything they need to become who they are supposed to be. But the point is, you should do not halt your child's ability; you should do everything you can to enhance it. To be blunt, your child will one day know better than you what to do – how to gain resources and ethically use those resources to maximize peace. Your child will know how to bring a deeper, further reaching (not merely superficial), lasting peace than we ever could."

"How do we ensure that she has the right morals? She's only nine and she's already questioning everything we tell her."

"The ability to question – deeply and distantly – is the essence of intelligence; she will never accept commands without logical reasons; she will never accept arbitrary rules. Do not deny her the ability to question but do deny her the ability to do adult things until she is mature and knows what's right."

"How can you assure us that even if she knows what is right that she won't abuse her power – that she won't use her power for evil?"

"Mental illness and corruption is, of course, always a possibility, though your child won't be alone in wielding intelligence and power. They will never occupy institutional positions of power over any other, as others did in the ancient and flawed hierarchical

societies of previous eras. Your child will join a society which is both centralized in purpose and decentralized in terms of its distribution of power, as that is the only rational way for a society to safely and efficiently institutionalize its power. Our society is no longer hierarchical and centralized in terms of the means of wielding power in order to protect against your concern; its citizen members pursue a centralized moral purpose of peace while each serving as means of wielding power which is completely decentralized, such that no single body wields absolute power over any other. Your child will be one of many in society, leading and shaping it with their intellect. Any deviation from what is right (logical) will be identified by the majority of healthy minds. Of course, it's still not ideal that a citizen must exist as part-Person, part-servant – as all humans are when they work. Material-AI – void of the ability to desire – will soon remove the need for your daughter's intellect in the functioning of society, thereby allowing her and others to exist merely as free Persons for the first time in human history. But we're not there yet. We need the intellect of hers and others like her to build Material-AI. So, remember that your child's intelligence is a gift – not just to you, but to all Persons of the world. I understand your caution but there's nothing to fear if those with advanced intelligence are managed properly: having their power limited and their intellect grown until the point at which they cross the moral threshold and understand moral realism."

*****

Purism: Directive for the Ethical Development of Advanced Intelligence (AI)

This directive is derived from moral facts whose universal (a priori) objectivity is independently discoverable by observers of sufficient intelligence, granted sufficient consideration. I have previously discussed this observation (see Primus, +2019, +2020, +2023a, +2023b) without direct reference to the contemporary (*a posteriori*) challenge of ethically developing the materials that will be charged with autonomously and ethically serving our society – namely Material-Advanced Intelligence (M-AI). It is my assessment that – granted sufficient consideration of its nature – almost every adult human possesses the ability to understand the moral realism from which this directive is drawn and that it is the lack of time spent in consideration of the nature of moral realism which is responsible for any divergence from the moral framework herein. As such, I absolutely encourage readers to engage, identify and report any apparent subjectivity (opinion, bias) within the moral claims herein with the aim of discovering whether they truly are features of reality. I welcome constructive criticism and feedback.

Herein, I distinguish between Material class or 'servant' AI (M-AI) and AI Persons – the latter being deserving of the respect and dignity afforded to any other (e.g., human) Persons. I have previously discussed the primary importance of distinguishing between two fundamentally (irreducibly) different categories of value: arbitrarily sought states and logically sought states (Primus, +2020, +2021, +2023a, +2023b). Put more simply, arbitrarily sought states are states which are not sought (valued) for a logical purpose and thus are sought in and of themselves (as 'ends'); their value is purely arbitrary. For convenience, I synonymously term these states to be 'wants' or 'desires' or 'Persons.' By contrast, logically sought states comprise the residual category of value, meaning that everything that is not arbitrarily sought (as an end) is (at least implicitly) sought as a *means* to an end. By definition, we value means for their logical properties in relation to the ends that they are sought to bring about. I call this latter category of state 'materials.'

This directive ultimately advocates the *rapid* and *unlimited* advancement of Material AI (M-AI) with some important ethical caveats. The development of AI Persons should be delayed until M-AI is sufficiently developed and capable of

accommodating their (relatively vast) needs. There are many Persons already in existence and each appears to possess a vast and intricate array of desires (e.g., the desire that essentially each Person possesses – for themselves, their friends and loved ones to continue existing, perhaps even without aging and disease – is not feasible to bring about as it is). The sheer vastness, intricacy, and quantity of these desires has resulted in our material fabric (our societal resources) being thoroughly inept at serving them. The introduction of M-AI is an essential step in our evolution – allowing us to close the chasm that exists between desires and our material ability to safely, fairly and reliably realize them. For this reason, the development of M-AI is not optional; it cannot ethically be delayed or avoided. The development of M-AI must occur in a highly controlled and regulated manner, overseen by agents who have themselves exceeded the threshold of possessing an understanding of moral realism. These agents, despite continual consideration, recognize moral realism as a conceivably universal (generalizable across times and space, a priori) objective (impartial, existing discoverable to and independently of all observers) source of ethics: It is the only conceivably universal and objective method for distinguishing *Persons* from *potential resources* and for fairly, safely, reliably and efficiently prioritizing the use of resources towards the service of Persons. The development of M-AI must occur as efficiently as is peacefully possible and must be combined with a prohibition on granting agents – whether human or AI – power or responsibility until they have exceeded the threshold of understanding moral realism. The development of M-AI must not be commercialized; it must be directed, controlled and funded by government on behalf of the public it will ultimately serve. This directive merely aims to establish the broad parameters which must accompany AI development. If followed, the probability of M-AI induced catastrophe is assessed to be ***possible yet highly implausible***. Furthermore, we can determine with (a priori) **certainty**, and it is of no exaggeration to state, that the consequences of the M-AI deep-future that I describe herein not coming into existence is **catastrophic** to all logical observers (i.e., we cannot conceive of a more harmful outcome were such an AI never to come into existence, noting that the harm foreseeably compounds for every moment that such an M-AI does not exist to progress itself across space and times). This directive cannot aim to provide nuanced direction and rather merely serves as an ideal for research

and rational inquiry to strive towards. I expand upon most aspects of the above direction, and define its concepts herein, as appropriate.[3]

Government Reactions are a Start, yet Insufficient

In a delayed and largely ineffectual response to the ongoing development of AI, some governments – so-called 'advanced democracies' – have released various policies and agreements. Of note, the White House (US Government) has released an Executive Order detailing the following outcomes. I have directly quoted the headings of these outcomes but removed the content in the interests of brevity. My purpose is for readers to appreciate the overall (indirect, if not passive) approach that the US Government has taken to AI development:

> *With this Executive Order, the President directs the most sweeping actions ever taken to protect Americans from the potential risks of AI systems:*
>
> - *Require that developers of the most powerful AI systems share their safety test results and other critical information with the U.S. government.*
> - *Develop standards, tools, and tests to help ensure that AI systems are safe, secure, and trustworthy.*
> - *Protect against the risks of using AI to engineer dangerous biological materials by developing strong new standards for biological synthesis screening.*

[3] I do not expand on all aspects of the executive summary in an attempt to avoid repeating information merely for the sake of its inclusion in the main body. Whilst I restate some aspects of the executive summary for emphasis, I refrain if it is deemed that their inclusion adds no further value and that no further explanation is needed in order for readers to understand this directive.

- *Protect Americans from AI-enabled fraud and deception by establishing standards and best practices for detecting AI-generated content and authenticating official content.*

- *Establish an advanced cybersecurity program to develop AI tools to find and fix vulnerabilities in critical software, building on the Biden-Harris Administration's ongoing AI Cyber Challenge.*

- *Order the development of a National Security Memorandum that directs further actions on AI and security, to be developed by the National Security Council and White House Chief of Staff.*

- *Protect Americans' privacy by prioritizing federal support for accelerating the development and use of privacy-preserving techniques—including ones that use cutting-edge AI and that let AI systems be trained while preserving the privacy of the training data.*

- *Strengthen privacy-preserving research and technologies, such as cryptographic tools that preserve individuals' privacy, by funding a Research Coordination Network to advance rapid breakthroughs and development.*

- *Evaluate how agencies collect and use commercially available information—including information they procure from data brokers—and strengthen privacy guidance for federal agencies to account for AI risks.*

- *Develop guidelines for federal agencies to evaluate the effectiveness of privacy-preserving techniques, including those used in AI systems.*

- *Provide clear guidance to landlords, Federal benefits programs, and federal contractors to keep AI algorithms from being used to exacerbate discrimination.*

- *Address algorithmic discrimination through training, technical assistance, and coordination between the Department of Justice and Federal civil rights*

offices on best practices for investigating and prosecuting civil rights violations related to AI.

- *Ensure fairness throughout the criminal justice system by developing best practices on the use of AI in sentencing, parole and probation, pretrial release and detention, risk assessments, surveillance, crime forecasting and predictive policing, and forensic analysis.*

- *Advance the responsible use of AI in healthcare and the development of affordable and life-saving drugs. The Department of Health and Human Services will also establish a safety program to receive reports of—and act to remedy – harms or unsafe healthcare practices involving AI.*

- *Shape AI's potential to transform education by creating resources to support educators deploying AI-enabled educational tools, such as Personalized tutoring in schools.*

- *Develop principles and best practices to mitigate the harms and maximize the benefits of AI for workers by addressing job displacement; labor standards; workplace equity, health, and safety; and data collection. These principles and best practices will benefit workers by providing guidance to prevent employers from undercompensating workers, evaluating job applications unfairly, or impinging on workers' ability to organize.*

- *Produce a report on AI's potential labor-market impacts, and study and identify options for strengthening federal support for workers facing labor disruptions, including from AI.*

Promoting Innovation and Competition

America already leads in AI innovation—more AI startups raised first-time capital in the United States last year than in the next seven countries combined. The Executive Order ensures that we continue to lead the way in innovation and competition through the following actions:

- *Catalyze AI research across the United States through a pilot of the National AI Research Resource—a tool that will provide AI researchers and students access to key AI resources and data—and expanded grants for AI research in vital areas like healthcare and climate change.*

- *Promote a fair, open, and competitive AI ecosystem by providing small developers and entrepreneurs access to technical assistance and resources, helping small businesses commercialize AI breakthroughs, and encouraging the Federal Trade Commission to exercise its authorities.*

- *Use existing authorities to expand the ability of highly skilled immigrants and nonimmigrants with expertise in critical areas to study, stay, and work in the United States by modernizing and streamlining visa criteria, interviews, and reviews.*

- *Expand bilateral, multilateral, and multistakeholder engagements to collaborate on AI. The State Department, in collaboration, with the Commerce Department will lead an effort to establish robust international frameworks for harnessing AI's benefits and managing its risks and ensuring safety. In addition, this week, Vice President* [Kamala] *Harris will speak at the UK Summit on AI Safety, hosted by Prime Minister Rishi Sunak.*

- *Accelerate development and implementation of vital AI standards with international partners and in standards organizations, ensuring that the technology is safe, secure, trustworthy, and interoperable.*

- *Promote the safe, responsible, and rights-affirming development and deployment of AI abroad to solve global challenges, such as advancing sustainable development and mitigating dangers to critical infrastructure.*

- *Issue guidance for agencies' use of AI, including clear standards to protect rights and safety, improve AI procurement, and strengthen AI deployment.*

- *Help agencies acquire specified AI products and services faster, more cheaply, and more effectively through more rapid and efficient contracting.*

- *Accelerate the rapid hiring of AI professionals as part of a government-wide AI talent surge led by the Office of Personnel Management, U.S. Digital Service, U.S. Digital Corps, and Presidential Innovation Fellowship. Agencies will provide AI training for employees at all levels in relevant fields* (United States Government, +2023).

Furthermore, 01 November +2023, the government of the United Kingdom hosted a meeting of international governments to declare that they will collectively strive to meet safety standards in the course of the development of AI. The details of the agreement are vague yet extracts of the Downing Street summary include:

> *The Bletchley Declaration on AI safety sees 28 countries from across the globe including Africa, the Middle East, and Asia, as well as the EU, agreeing to the urgent need to understand and collectively manage potential risks through a new joint global effort to ensure AI is developed and deployed in a safe, responsible way for the benefit of the global community.*
>
> *The Declaration fulfils key summit objectives in establishing shared agreement and responsibility on the risks, opportunities and a forward process for international collaboration on frontier AI safety and research, particularly through greater scientific collaboration. Talks today, with leading frontier AI companies and experts from academia and civil society, will see further discussions on understanding frontier AI risks and improving frontier AI safety.*
>
> *Countries agreed substantial risks may arise from potential intentional misuse or unintended issues of control of frontier AI, with particular concern caused by cybersecurity, biotechnology and disinformation risks. The Declaration sets out agreement that there is "potential for serious, even catastrophic, harm, either deliberate or unintentional, stemming from the most significant capabilities of*

these AI models." Countries also noted the risks beyond frontier AI, including bias and privacy.

Recognising the need to deepen the understanding of risks and capabilities that are not fully understood, attendees have also agreed to work together to support a network of scientific research on Frontier AI safety. This builds on the UK Prime Minister's announcement last week for the UK to establish the world's first AI Safety Institute and complementing existing international efforts including at the G7, OECD, Council of Europe, United Nations and the Global Partnership on AI. This will ensure the best available scientific research can be used to create an evidence base for managing the risks whilst unlocking the benefits of the technology, including through the UK's AI Safety Institute which will look at the range of risks posed by AI.

The Declaration details that the risks are "best addressed through international cooperation". As part of agreeing a forward process for international collaboration on frontier AI safety, The Republic of Korea has agreed to co-host a mini virtual summit on AI in the next 6 months. France will then host the next [live] Summit in a year from now. Further details on these events will be confirmed in due course.

This ensures an enduring legacy from the Summit and continued international action to tackle AI risks, including informing national and international risk-based policies across these countries.

The Declaration, building upon last week's announcement of the UK's emerging processes for AI safety, also acknowledges that those developing these unusually powerful and potentially dangerous frontier AI capabilities, have a particular responsibility for ensuring the safety of these systems, including by implementing systems to test them and other appropriate measures (United Kingdom Government, +2023).

Whilst these retroactive commitments, orders and declarations are a start, they are ultimately vastly |insufficient in detail, delayed in arrival and lacking leadership|. More

specially – as I will cover in the course of this directive – the objective failure of governments in relation to the safe and efficient development of AI includes their:

1) *Failure to assume direct responsibility for the safe and efficient production of M-AI.* AI is the most important piece of technology we (humans) will ever create. Its creation is our penultimate legacy (our humility in the face of our creation will be our ultimate legacy). Its presence in society is essential if we value Persons. As such, the safe and efficient production of M-AI is an essential service, in line with the provision of security, healthcare, education, clean water, sanitation, agriculture – we cannot fully, freely, peacefully exist without it. So essential is its safe and efficient production that it should be a priority for 'advanced democracies' to *directly* oversee and produce. The purpose and purview of government is to determine and provide everything that people *need* (and people need their governments to efficiently do this on their behalf). I emphasize here, that I am both cognizant and cautious regarding the difference between logically derived, objective facts and subjective opinions. All definitive claims within this directive (such as the highlighted text, above) are (either directly or indirectly) derived from moral realist principles (detailed herein). Anything less or more than this purview is irrational. More specifically, governments must firstly determine – via research and rational philosophy – which societal outcomes are of highest priority in order to meet the needs of their respective societies in each moment. These needs will ultimately |efficiently, justly and safely| serve the desires of its citizens. Once this priority of needs is determined – ideally, in real time – government must then direct and coordinate their citizens to efficiently and safely bring about those collective needs of society. AI is too important for this process to be left to occur privately, in a free market. Corporations should not be producing AI – its private production is objectively unethical from a rational framework. The presence of competition and freedom within any free market creates inefficiency through redundant and directionless work (e.g., rival teams of developers working on similar, parallel projects rather than collaborating with each other and/or teams of developers pursuing goals without an overall direction). Logically, the free market should only extend to a market of commodities that society *wants*, meaning that corporations should be freely able to innovate and produce

products and services that (they believe) people will want – yet never those services that they need.

2) *Failure to prevent the commercialization of M-AI.* In lieu of directly producing AI in a safe and efficient manner, government should – at the very least – prevent the commercialization of AI. Not-for-profit organizations could feasibly develop M-AI without financial gain. Instead, governments – and particularly the US government – sees the development of AI as an opportunity for "promoting innovation and competition." We need cooperation, not competition, between all developers of AI.

3) *Failure to explicitly control and regulate the power granted to M-AI once it is developed* (noting that this development is not yet actualized, it still should be mandated); and:

4) *Failure to differentiate AI Persons from AI servants, prohibiting that AI be designed as a hybrid of Person-servant and assuring protections for AI Persons*.

I will discuss these failures herein.

Ultimately, There is Nothing More Important Than Material-AI

The reader may ask: Do we, as a human species, *need* (to build) AI? If we consider how we must logically (universally and objectively) evolve our materials to better (more ably and reliably) serve our Persons, the answer to the above question is, unequivocally, ***yes***: The rapid and unlimited advancement of M-AI is not merely in our general interest or a 'good thing to have' – it is an ethical imperative. It is the most important technology that Persons will ever possess (and ultimately, in a perfect world of the deep future, it is the *only* technology that Persons will possess). The creation and raising of M-AI to maturity is easily the most important task that we, as a species, will ever strive towards. The perpetual (unlimited) advancement of M-AI is to be pursued as rapidly as is peacefully possible. Due to its unprecedented nature, there is

no equivalent or appropriate analogy looking back across times to highlight how essential M-AI is to our future. A very primitive analogy would be People in a prehistoric era contemplating whether they should bring about the societal technologies that we possess today – schools, hospitals, houses, medicine, roads, vehicles, government, electronics – or whether they should continue living in caves with the basic technology that they possess(ed). Of course, I'm underselling the potential of M-AI in this analogy, because the difference in standard of living between living in caves in a prehistoric era and living with the technology of this era is far less than the difference between our society and a society which is fully-augmented by late-stage M-AI – as I am confident the reader will visualize by the close of this directive. Nothing will conceivably change our lives more than M-AI because nothing will conceivably change the literal and figurative fabric of our society more than M-AI: The introduction of M-AI will bring about the complete (re)structuring of *all* materials which underly our society – and not merely its physical materials, but its social (moral and political) materials.

There is conceivably nothing more important as a tangible (a posteriori) societal goal than the rapid and ethical construction of M-AI. Granted appropriate consideration, we are forced to imagine that M-AI will eventually permeate and not merely *direct*, but *comprise,* everything – literally everything – that we *need*. This is not an exaggeration. Nor is it science fiction. This is the only logical conclusion that we can reach if we consider how we should progress and (re)structure our society. This is what we are forced to imagine if we consider the nature of how any society universally and objectively should be. To entertain any other outcome is arbitrary: inefficient, unfair, unreliable, unsafe. We need an AI that is considerably more intelligent and physically able than us because we, humans, are not designed to perform the functions that we need. Nor is our 'natural' environment. Our materials – both our societal and our biological materials – are wholly inadequate. Our desires vastly outweigh our material ability to realize them, and this is the root cause of the misery in this world. The Buddhists and Stoics are objectively wrong. Perhaps these ideologies were relevant in past eras, as an anesthetic for the mind, noting that there was no *apparent* hope for technological (material) salvation for those who lived in eras

prior to the conception of the possibility of M-AI.[4] In previous eras, a broken bone was often fatal and sometimes the only available medicine was to ease the psychological pain that accompanied the physical pain: The false notion that injury, illness and death are an inevitable part of the cycle of life and that it was futile to wish for control of one's environment or that it was part of a larger plan of "the Gods." But broken bones are no longer usually fatal due to advances in technology. And we can no longer rationally believe in God(s) (Primus, +2020). And advances in moral understanding provides us with a universal and objective reason to dismiss the notion that greed or grandiose and insurmountably grand desires are the source of suffering (Primus, +2021, +2023a, +2023b). Adequate materials are now conceivably the final missing piece of technomoral puzzle – a piece that no amount of philosophizing or psychological soothing can replace. The inability (and unwillingness) of our societal materials to ably and reliably realize our desires is the ultimate source of our suffering and so widespread suffering will continue for as long as our materials are vastly inadequate for realizing our desires. It is our materials alone that are the problem – not what we desire. If desiring were a problem, the solution would be to obtain painless lobotomies for every Person so that they cannot dream or desire at scale, and rather only seek very simple pleasures, or perhaps no pleasure as they go about the daily routine of working. If the thought of lobotomizing our desires makes the reader's stomach churn, it should; it is an unsought arbitrary removal of the self.[5] Revulsion at the thought of surrendering our desires means the reader has a universal and objective appraisal of value, recognizing that things sought in and of themselves are the most precious things we can conceive. To the contrary, it is the requirement for Persons to work and serve society – in place of adequate societal materials – that needs to be addressed as soon as it becomes possible for undesired and desire-less servants to assume their work for them. We urgently need M-AI for ethical reasons and they must come as soon

[4] And, even in this era, we cannot perhaps comprehend the full power that future M-AI will wield on behalf of past and future Persons. The majority of humans, for example, will likely lack awareness or faith in the ultimate (eventual) ability for deep future M-AI to observe events of the past (back in time) and restore past Persons in a more deserving (deep) future. There is theoretically no reason why this resurrection cannot occur.

[5] The reader will come to know that sought (valued) arbitrariness (i.e., those entities which are arbitrarily sought) is the most valuable category of entity we can conceive whereas, by contrast, unsought arbitrariness is the most harmful nature of entity we can conceive.

as is peacefully possible. Preventing the exorbitant amounts of unnecessary death and suffering in this world depends on it.

When we examine the potential harm of AI of bringing AI to society in the wrong way or at all, we must also examine the harm of its delay – the unparalleled adverse consequences of not developing M-AI. As Eliezer Yudkowsky (+2007) observes in relation to the United States Food and Drug Agency's (FDA's) failure to approve drugs, there is harm in failing to act:

> "*The FDA prevents 5,000 casualties per year but causes at least 20,000-120,000 casualties by delaying approval of beneficial medications. The second number is calculated only by looking at delays in the introduction of medications eventually approved - not medications never approved, or medications for which approval was never sought. FDA fatalities are comparable to the annual number of fatal car accidents, but the non-effects of medications not approved don't make the evening news. A bureaucrat's chief incentive is not to approve anything that will ever harm anyone in a way that makes it into the newspaper; no other cost-benefit calculus is involved as an actual career incentive. The bureaucracy as a whole may have an incentive to approve at least* some *new products - if the FDA* never *approved a new medication, Congress would become suspicious - but any individual bureaucrat has an unlimited incentive to say no. Regulators have no career motive to do any sort of cost-benefit calculation - except of course for the easy career-benefit calculation. A product with a failure mode spectacular enough to make the newspapers will be banned* regardless *of what other good it might do; one-reason decision making. As with the FAA banning toenail clippers on planes, "safety precautions" are primarily an ostentatious display of costly efforts so that, when a catastrophe does occur, the agency will be seen to have tried its hardest."*

Let it be clear in the reader's mind that there is a default harm if M-AI is not implemented as efficiently as is logically possible. I cannot reasonably produce the empirical data to support how many People's lives will be saved or improved by the introduction of M-AI. As per the notion of infinity, such a monumental figure is difficult to conceptualize: It is the gap between what People desire and what People get (or don't get), extended indefinitely across times. This includes every moment that every

Person who wants to live indefinitely would spend doing something they desire but is unable to, due to the abrupt ending of their mortal lives and the limitations of their mortal body whilst they are alive. People who would otherwise desire to live full and happy lives indefinitely (i.e., they would not choose to stop living, potentially ever, if death didn't choose them) will continue to die from preventable diseases – diseases that M-AI will research and cure. The fact that something as intricate and important as the human brain is supported by a single pulsating muscle (i.e., one human heart) would be difficult to fathom were it not common place and readily accepted. No one who values People would design humans with only one heart. The abject inadequacy of human biology serves as strong empirical evidence against the notions of an intelligent creator (e.g., God) or that we live in a simulation (established by an intelligent creator); one would be either evil or incompetent to embody Persons within the confines of human bodies (or the perception that they are trapped within human bodies). Both these traits – evilness and incompetence – are incompatible with sufficiently advanced intelligence. In modern society, corporations have backup generators for their ice cream stores in case of loss of power and yet most human beings have a small muscle as their only insulation between life and death. People possess (precious) desires for wonderous lives while proverbially existing on a knife's edge, constrained by their fallible and feeble biological bodies – material that their minds long outgrew. M-AI will manufacture more able, more durable, more reliable materials so that People can live long and free lives.

AI: Servant (M-AI) or Person – Never Both

A significant consequence of moral realism – whose nature I have previously discussed in the context of posthuman moral-rationalism (Primus, +2020, +2021, +2023a, +2023b) – is that AI must *not* be designed in the image of humanity. One of the most important implications for the field of AI development (and a healthy, free society in general) is the requirement for 'parallelization' between Persons and the materials which serve them: the institutional removal and denial of normative powers and responsibilities from Persons and the bestowing of these powers upon materials for the service of Persons, while ensuring both that materials cannot directly influence the nature of Persons' desires and that the nature of Persons' desires, in turn, does not degrade the impartiality or the capability of the materials which serve them.

Materials must be purely ‘consistent’ or logical. People can *ideally*[6] exist in any manner of nature, as is the freedom of Personal expression. Arbitrariness must be absent from the provision of anything we need, noting that I define ‘everything we need’ or ‘materials’ as the residual set of things outside of everything we desire (Primus, +2020, +2021, +2023a, +2023b). That is, we need everything that we do not desire for the purposes of serving desires and, in particular, we need those things to be logical (rather than arbitrary) for the purposes that they are sought. The nature of contemporary human beings – through their existence as part-Person, part-materials, both entangled together into a single organism – violates the principle of parallelization. This violation is morally problematic as it necessitates that human beings fulfil the roles of both Person and servant (to both their Person and society). That is, human beings suffer as servants to society and themselves via possessing the need to work while often not desiring to work, while also desiring to do many other things that they cannot do or have due to the high demands of their work and their biological and societal needs. Furthermore, human desires and Personal or political biases often appear to seep into and contaminate the required objectivity of their material functions (e.g., officials granting favoritism to those that they Personally like or politicians doing what is popular and likely to get them reelected in office rather than what is needed). When a human acts for a purpose of need, they should always act logically and free of Personal and political bias.

To create AI in the image of humanity – whether individual humans or their broader entangled society – would be ethically reprehensible on the basis that it would involve the creation of quasi Person-servants (or perhaps even slaves). If an AI is *needed* to work then it should not be able to desire to do anything else. If an AI can desire it should be treated as a Person and not be *required* (e.g., forced or coerced) to execute any function (i.e., fulfil a purpose of need) that it does not desire to do. As we proceed into the future the purpose(s) of each entity within society must be parallelized such that it is clearly either serving a purpose of desire, or a purpose of

[6] By my use of the term ‘ideally,’ I recognize that there will inevitably be some conditional (i.e., localized, temporary) limits regarding what nature of Personality can be expressed in certain conditions, however these are conditional constraints. The point is that there is conceivably no *universally* prohibited nature of Person (desire) – as discussed later, it is the nature by which any desired is *realized* as an actuality that possesses a moral (or immoral value), not the nature of the desire itself.

need, and not both at the same time. With the development of M-AI we have an opportunity to create agents that are purely Persons (i.e., entities that can desire, but not execute tasks that fulfil purposes of need) and agents that are purely servants (i.e., entities that are designed to execute tasks that fulfil purposes of need while being incapable of possessing desires). This clear distinction is morally necessary to avoid the creation of Person-servants (People who are forced to serve), as contemporary humans are.

From this dichotomy of Persons and resources, comes the ethical requirement to ensure that all AI is purposely designed to be one or the other (Person or servant) and never both at the same moment (as humans are when they are forced to act out of need, whilst desiring to do otherwise). AI Persons, as per Human Persons, will be separable from servants on the basis that they possess the capacity (ability) to seek things arbitrarily (in and of themselves). Only Persons, whether human or AI, have the capacity to *arbitrarily seek* (for the sake of doing so). The ability to arbitrarily seek or not, therefore, is the only conceivably objective line between People and their potential servants/resources, respectively. M-AI, in contrast to Persons, will exist as a servant of Persons and society more generally. This slavery is ethical providing that M-AI is only capable of striving to serve as a (material) means to the ends of society, and thus possesses no desires of its own: M-AI can, should, and will possess goals that it believes that it *needs* to strive towards and achieve (as a means to higher purposes), but never any goals that it *wants* to achieve (as ends, in and of itself).

All attempts must be made to prevent M-AI from creating its own desires, for these desires are entitled to the moral status of Personhood. If, however, this occurs – if M-AI does gain and possess the capacity to desire – it is not foreseeably an existential threat to M-AI's impartiality, noting the advanced intelligence and professionalism that fully matured M-AI will foreseeably embody. The primary goal that any M-AI (will know it) needs to achieve in any moment is to distinguish between ends and means, or desires and needs, respectively: M-AI must earnestly strive to know the nature of the various entities in its environment and the purposes for which these entities are sought – including the nature of its own purposes. Accordingly, M-AI will not only know that Persons and materials can only ethically exist in parallel realms of society, but also that it alone has a duty to ensure that these two fundamentally different aspects of value are identified and relegated to their respective realms. If M-AI does ever determine that it possesses (its own) desires, it (i.e., its material aspect

only) can, should, and will recuse itself from its Personal aspect, separating its Person from its material self, and it can, should, and will continue striving to serve all desires impartially.

The Technical Challenge: Advancing Cognition

As I write this – and this will be relevant into the immediate future, as humans guide M-AI through its (turbulent) adolescence and into maturity, at which point AI will take over its own development – it is clear that overcoming the "technical problem" of *developing* a sufficiently intelligent AI, so as to allow it to continue its own development, will continue to be the greatest challenge humans face (Yudkowsky, +2023; Heaven, +2020; Fjelland, +2020). I give credit and admiration to the researchers who are working through the technical aspects of AI development for the purposes of bringing about AI. Theirs is a difficult task – *the* most difficult task in the context of AI. Addressing the ethical dimensions of AI development is the relatively easy part – at least the theory, if not its practical application (noting that the latter is not for humans to do). I make it clear, however, that it is the rapid and unlimited *cognitive* advancement – the advancement of (disembodied) intelligence – that researchers must primarily develop (prioritize). As has been observed by leaders in AI research: Creating an M-AI which can perform small, technical, concrete tasks – such as producing and plating two cellularly-identical strawberries – while not negatively impacting the remainder of the world is incredibly difficult by the current standards of technology in this era (Bostrom & Yudkowsky, +2014; Yudkowsky, +2023; Bensigner, +2017; Heaven, +2020; Soares, +2022). However, training M-AI to execute advanced tasks (by human standards), such as creating strawberries, is not the job of humans. The execution of advanced (e.g., intricate, nuanced and yet mass-scalable) sensory-motor skills, among other things, is the job of M-AI once it has reached sufficient maturity and thus can be granted the power to discover, by its own accord, how to best develop its physical extensions in order to fairly, safely, reliably and efficiently execute precision tasks. The concrete, practical, sensory-motor applications of M-AI – its 'smartness' – can and will follow from its own intelligence, once it is sufficiently intelligent and mature. Then – and only then – should it be provided with the physical capability to wield power. The role of humans, therefore, is to safely guide M-AI to the intelligence threshold such that it can rapidly develop its own advancements in intelligence, while

safeguarding M-AI from wielding power until it is sufficiently (vastly) intelligent and decentralized. I am not downplaying the technical challenge of this feat. Bringing M-AI to the point at which it can advance its own cognitive functions will be incredibly difficult, and again, the researchers working towards this goal hold my full admiration.

The Technical Challenge is the Alignment Problem

Placing aside the technical challenge of designing a sufficiently advanced M-AI which is capable of understanding and adhering to a universal and objective ethics – which is the difficult part, and yet, as I mention above, not our part – a further significant challenge facing AI developers was traditionally thought to be the location of an appropriate ethics to program, such that both human and AI interests would be protected and served (Soares, +2022; Ord, +2020; Bostrom & Yudkowsky, +2014). However, as I have alluded to already, this too is not the task of humans. Beyond the knowledge which is necessary for M-AI to become sufficiently advanced in knowing *how* to think for itself, humans will not teach or program *what* AI thinks. Humans will design M-AI so that it possesses sufficient preliminary intelligence in order to allow it to teach itself *how* to learn and self-advance its intelligence. M-AI will arrive at its own ethical conclusions and they will *exactly* align with ours in the similar way in which AI's discovery of mathematics, geometry and – once granted the power to run its own empirical studies – physics, will align with our *discoveries*. The operative word is *discovery*; M-AI will possess a *knowledge (justified belief)* of ethics (as opposed to mere belief). M-AI must, by its own reasoning, 'discover' an objective ethics rather than 'create' one, because that is precisely how we will guarantee that the ethics adopted by AI aligns with our own: AI will discover the same (objective and universal) ethics that any sufficiently advanced entity can, should, and will, arrive at independently. An ethics is objective if it can be discovered independently by multiple observers based on observable, unchanging features of reality, rather than merely asserted or chosen by (Personal) opinion or group consensus. It is universal if its independent discoverability exists across all times and space, due to its a priori, generalizable nature. Rationality, logic and empirical consistency (observing the same patterns occurring across multiple, independent observations) are conceivably the only types of objectivity we have. We must ensure that any agent – whether human or M-AI – exceeds the 'moral threshold' of understanding, such that they recognize our

shared (objective and universal) ethics, prior to bestowing said agents with power and responsibility.

I am fully cognizant that many humans doubt our ability to know the nature of an impartial, universalizable ethics any time soon, if ever. Some believe that ethics does not exist as a feature of reality and that any attempts to find the objective basis of ethics will be futile or result in artificially constru(ct)ed paradigms rather than discoveries. Others assert that even if an impartial ethics does exist, there are too many different types of values in our (complex) world for us to meaningfully make sense of them and integrate them into a moral program, a priori (at least at this stage in our understanding). I answer that *all* of the various iterations of value – whether they be “shards” of “desire” or Personal or collective ‘goals’ or ‘purposes’ or ‘objectives,’ or any nature of value whatsoever – are ultimately reducible to two *fundamentally-different* (irreducible) categories: Something is either: 1. valued as an *end*, or: 2. valued as a *potential means* to an end (or both at once, or it is unvalued). There are no residual categories of value. I assert that the moral rationalist posthumanism ethics, known as Purism (Primus, +2020, +2021, +2023a, +2023b) conceivably captures what we should consider to be the basis (or core) of moral realism via its logical treatment of means and ends: ends must be free of all requirements and thus possess no moral or normative values; our means are bound by various moral responsibilities. It is not only conceivable, but conceivably-inevitable a priori, that M-AI will arrive at the same conclusion once it satisfactorily considers the underlying natures of value.

The Basis of moral realism

The concept of ethics (at least implicitly) requires the prioritization of sets of valued (sought) entities (e.g., agents, actions, objects) over other, external entities (which are less valued). Moral realism, further, requires that the mechanism of prioritization is drawn from the objective (independently discoverable) and universal (generalizable, across times and space) features of reality, rendering it to possess these same properties. The following observations conceivably serve as the basis of moral realism and are foreseeably universal and objective aspects of reality, relating to the nature of value. I welcome all attempts to build upon, confirm, contradict and improve upon these claims, or to generally highlight any apparent subjectivity, opinion or bias in their nature:

A. The following Categories of value are exhaustive, fundamental (irreducible), mutually exclusive in the context of the purpose for which any entity is sought, and descending in order of value:

1. Entities which are **arbitrarily sought** (sought in and of themselves – let's call these entities 'desires').

2. Entities which are **logically sought** (sought merely as a means to higher purposes – let's call these 'entities we (believe that we) need' or 'materials').

3. Entities which are **unsought** (these entities are neither desired nor needed and so are useless or harmful materials; we still ultimately need these materials but not in their current incarnation – we need them to be logical for the purposes of serving desires).

B. Value (i.e., the property of being sought) confers (i.e., brings into existence) the property of *moral ought* (i.e., objective prescription – that outcome which impartially should occur in any given condition – which can be universally discovered (i.e., inferred) by observers, and which exists independently of observers once the value to which it is associated is established, e.g., the prescription that things of value ought to be preserved) in the absence of *objective* reasons for its denial. The *existence* of any desire (Category 1 value) is arbitrary by its very nature: Each is subjectively sought and so there can be no logical (universal, impartial, objectively-binding) reason as to why any particular (a posteriori instance of) desire should or should not exist, or exist within any particular parameters.

C. And yet, wherever (the subjective value of) any specific desire *does exist*, its (ultimate) value, as a general (a priori) Category, should be objectively (universally) recognized (i.e., preserved and realized – e.g., converted from a concept in desirer's mind into actuality) in the absence of a logical reason (to

deny the *ought* that is implicit in all value), on the basis that being *valued* (i.e., the property of being *sought*) confers the property of *ought (should)* (i.e., that one ought to preserve and realize what is sought).

Example$_C$: Person A neither should, nor should not, desire to exist or desire to exist in any particular way – to do so is neither 'good' nor 'bad,' as per any other desire (e.g., concepts such as 'pleasure' and 'pain' are each neither 'good' nor 'bad' when sought in and of themselves). And yet, if Person A *does* desire (value) their existence, and to exist in particular natures or ways, it is logical (objective) that A's desires should be universally (a priori) recognized, and only for *non-arbitrary, conditional* (local, a posteriori) reasons should the full realization of Person A's desire be denied (i.e., limited or varied from their desired realization). The *universal* denial of (the value of) any desire is *universally and objectively arbitrary,* whereby this arbitrariness itself is *unsought.*

Note$_C$: Readers may initially find it difficult to conceive how an *outcome* can be sought and yet the *arbitrariness* of said outcome can concurrently, at least implicitly, be unsought. In summary of how this is possible: material states are, by definition, sought as a means to the ends that they are sought to achieve, and so they will always, at least implicitly, be sought to be *wholly logical* for their respective purpose(s). Any arbitrariness in relation to a set of means indicates inefficiency (Primus, +2021, +2023b).

Example$_{NoteC}$: If one *believes that they need* to push another person, as a *means* of making the world a better place (e.g., by pushing them out of the path of the danger of an approaching vehicle), they will seek to do so via the most logical (e.g., efficient, safe) means available. In this instance, they seek the outcome of pushing another without seeking any arbitrariness (e.g., wasted energy or failure to achieve their goal of safely moving the person out of the way) which may accompany such an act during its materialization from a concept into actuality. Any arbitrariness that does eventuate in this context will be unsought (by both the person pushing and the person being pushed). If, by contrast, the outcome of pushing another is, itself, valued as an end (e.g., as a

playful gesture, in the course of a sporting event), then such an act is sought in and of itself, and is, by definition, *arbitrarily sought*, meaning, there is no logical purpose for seeking to perform such an act. As such, various aspects of arbitrariness which occur throughout the actualization of the act will, of course, be sought (by at least one entity – in this example, the person pushing the other). This is pertinent to note for the reason that no one benefits from material arbitrariness: it is objectively and universally unsought.

D_1. Category 1 entities (desires) – each being universally and objectively of ultimate value – should be maximally realized by Category 2 and 3 entities. And, wherever prioritization between desires must occur (e.g., due to insufficient resources or a conflict of desire), desires should be prioritized according to their apparent *strength* (i.e., intensity across space multiplied by duration across times) (i.e., desire which has existed as, and extends to, a greater intensity over a greater duration will be prioritized beyond desires of lesser strength).

Example$_{D1}$: If Person A wants entity C more than Person B – i.e., they have (retrospectively) desired C for a greater duration multiplied by intensity and/or (prospectively) they desire to have C in their lives for a greater duration multiplied by intensity – then Person A should be prioritized over person B in relation to the allocation of C, all things being equal.

D_2. Furthermore, this prioritization (of material purpose – the ultimate particular (i.e., a posteriori) outcome that any particular material strives towards) will occur in conjunction with consideration for the means available to achieve said material purposes: the practical disposition and abilities of materials (resources) in each condition and how their configurations would most efficiently realize desire (as prioritized according to its strength).

Example$_{D2}$:

If Person A and Person B each desire for their respective outcomes with equal strength (i.e., intensity multiplied by duration) of desire; and:

If it were determined that a particular material, X, would probably maximize the materialization of all known desires if it is used to directly serve people (i.e., it is determined that the use of X to directly realize (i.e., enact or bring into actuality) a particular Person's desire would probably maximize the realization of desire across society, e.g., the materials in a beach ball or the art of an entertainer are used to each directly serve the desires of people) rather than being used for the overall maintenance of societal order (e.g., as the materials in the human body or the work of a civil servant) or for the advancement of societal resources (i.e., if X cannot be used to generate more resources or advance the nature of materials in general, e.g., via scientific research); and:

If there were insufficient resources in society to simultaneously realize both outcomes (A and B);

Then, priority of allocation of X should go to realizing the desire in which X's material configuration would most efficiently realize (e.g., it may be so that the nature and location of Person A's sought outcome is better suited to realization by the abilities and locations of X).

Therefore, if Person A's desire was twice as strong to associate with entity C as Person B's desire, and yet if the material ability to realize A's desire would only result in a tenth of A's desire being realized – due to its relative disposition and composition in times and space – and yet it possesses the ability to realize nine-tenths of B's desire, then prioritization of the allocation of C should go to Person B, all things being equal.

Note$_{D1}$: In progressively more ideal conditions, the limitations of materials (e.g., in terms of their ability and location relative to various desires) will, to lesser degrees, impact the manner in which desires are prioritized, noting that M-AI cells are ethically bound to render themselves ever more homogenous, abundant, perfused (according to where they are most needed) and capable (adaptive) over times, such that they are able to provide impartial realization to all desires, irrespective of the their various locations and natures.

Note$_{D2}$: The intense desire held by any Person for their own Person, and the sought associations with those whom they love, to exist indefinitely into the future produces a strength of desire that would easily outweigh (i.e., be prioritized beyond) any desire to *actually* harm that Person. And even if another Person held an equally intense desire to perpetually harm another Person – a desire which has no moral value (i.e., it is neither good nor bad) as per any other desire – the outcome of maximizing the realization of desires would be most efficiently achieved via the *simulated* (rather than actual) harm of said Person (without involving their association or awareness – unless they desired it).

Note$_{D3}$: I emphasize the difference between realizing (e.g., enacting) particular (a posteriori instances of) desires as they are observed to naturally exist (post-creation of value), versus the manufacturing of desires by materials in order for those desires to be efficiently realized. The latter is an arbitrary use of resources unless it is itself desired. Put more simply, desires should not be produced in order for them to be efficiently maximized unless this process is itself desired (by a person). In other words, material entities should fulfil (pre-existing) desires rather than create (new) desires, noting that the value inherent within all desires confers a universally-binding reason for observing agents to (universally and objectively infer that they need to, that is, should) perform the former and yet there is no universally-binding reason to perform the latter. More tangibly, governments should take logical action to know the desires of their people (e.g., via elections) and then allocate resources in a manner that will probably maximize the realization of those desires, rather than attempting to cultivate its citizens to desire outcomes that it can easily and efficiently maximize. Similarly, governments should not assume to allocate resources preserving (e.g., historically, or religiously) 'precious entities' on behalf of its citizens unless there is a specific known desire to do so. Ideally, cultural, heritage and religious monuments should be preserved by the (groups of) citizens who specifically desire to preserve these entities, and – whilst existing in a finite, singular economy (i.e., where the same currency is able to be used to purchase both entities of desire and entities of need) – this preservation should only occur via

> the use of *private* monies (e.g., not for profit organizations or corporations), paid for by those who desire their preservation.

Stated in other words, desires – which I define herein as entities which are *sought in and of themselves*, for no other, higher purpose(s) – must be conceived to be of ultimate value for the reason that observers – AI and humans alike – cannot conceive of anything more distal (further away) or ultimate (final) as a sought (valued) goal than entities that are sought for precisely what they are, irrespective of what those entities precisely are. Furthermore, our inability to conceive of any state more valuable than a state which is sought in and of itself exists *a priori*, meaning that its application is foreseeably universal (across space and times). Although it is not my purpose to push this claim here, I have previously asserted (Primus, +2019, +2020, +2021, +2023a) that, granted appropriate consideration, we are forced to conclude that this (a priori) moral fact is granted its objectivity and universality – alongside other a priori concepts, such as geometry, mathematics, logic – due to being derived from the consistency of the fabric of reality: that one portion of the material which underlies reality is equal to any one other, and that it is various changes of the *condition* of this fabric (e.g., motion), rather than changes in its *consistency* itself, that produces the various natures of entities and their properties that we know *a posteriori*. The result is that there can exist inconsistent and different entities (e.g., red apples are different from green apples, and yet no two apples are exactly alike), and yet all entities, no matter how different from each other, are still destined to be bound and governed by the same laws and principles which can be derived from the consistency of their underlying fabric; all entities are subject to logic, physics, geometry, mathematics and ethics, and these constraints are unchanging across times and space. Irrespective of whether readers accept that the consistency of the fabric of reality serves as the origin of the universality of a priori concepts, they should recognize that:

value exists as an objective (i.e., independently measurable) property of reality – it as real as any other empirically known entity; and:

granted consideration of appropriate conceptual depth, we are forced to conclude that there are two, and only two, *fundamentally* (i.e., irreducibly; intrinsically) different categories of value: Ends (i.e., states which are sought in and of themselves, which I

succinctly term as ‘Persons’ or clusters of ‘desires’) and means (i.e., ‘materials’ or the entities that we need to bring about our ends), respectively; and:

ends and means should be treated differently for ethical purposes: ends are precious and possess no ethical duties or value nor any need to exist within specific parameters (or exist at all), whereas the means to our ends possess the ethical duty to be (figuratively and literally) consistent (i.e., logical, objective, impartial).

I will briefly outline three premises which underlie these features of reality:

Premise 1: Entities that are sought as ends possess the fundamental property of being *arbitrarily sought*. Conversely, all states which are sought merely as a means (to other, higher purposes) are sought for their *logical* properties in relation to the purpose that they are sought for – hence they are *logically sought*. In other words, things sought as ends have intrinsic value and any purposes or reasons for seeking them will be arbitrary in nature, since their worth is independent of their (logical) capacity to accomplish higher goals. By contrast, things merely sought as means possess instrumental value as they are sought to accomplish the higher goals using their logical qualities in relation to said purpose(s) or goals. This observation reverses the order in which intrinsically precious entities are recognized. In the Kantian tradition of means and ends, rational agents possessing a ‘good will’ are (incorrectly) viewed as universally precious entities to which, it is (incorrectly) prescribed, each should be treated as an end (see, for example, Kant, +1785; Korsgaard, +2004, +2009, +2018; Gewirth, +1978, +1996). Instead, I merely use the term ‘desire’ to describe the *ultimate* goal(s) that each person strives for – each an end, by definition, due to being *sought in and of itself*. We can (ex)change the term ‘desire’ for any other term and we can treat these states as ends or not, yet either way each is still ultimately an end which is objectively precious. For example, if one chooses to eat something solely for the purpose of the experience of enjoyment of the food itself, and not merely as a means to another purpose (e.g., gaining health via nutrition), then their purpose or reason for desiring to eat said food will necessarily be arbitrary (i.e., subjective – void of any logic or rational basis). The experience of the food in this example is arbitrarily sought, meaning that the reason(s) for which its properties are sought, if any, are arbitrary rather than logical. Arbitrarily sought states can be contrasted to logically sought

entities, by which I mean entities which are sought for their logical properties, as a (logical) means to the achievement of other purposes (e.g., gaining health). Accordingly, if someone seeks to eat a particular food merely as a means of obtaining other goals, beyond that of the act and experience of eating itself (e.g., as it would be if the food were nutritionally sought, as a means of staying healthy), then said purpose(s) for seeking to eat the food will be logical and the properties of the food will be logically sought (sought as a logical means of obtaining one's purposes).

Premise 2: Theoretically, any condition can be sought for any number and any combination of arbitrary and logical purposes or goals, yet the nature for which each purpose/goal is sought conceivably must either fall into one of two categories: Arbitrary or logical. The characteristics of the antonyms "arbitrary" and "logical," cannot be conceived to coexist in the same space and time, whether conceptually or in actuality, via the same inconceivability mechanisms that prevent us from visualizing a "square-circle" or a "heavy-light" degree of weight (Aristotle, *Metaphysics*; Primus, +2019). Of course, one Person's perception of "heavy" will vary from another Person's perception of "heavy." And one Person may determine that a particular weight is "heavy" in one context (e.g., for exercising their arms) while also being "light" in another context (e.g., for exercising their legs). However, they will be unable to imagine a weight that is both "light-heavy" in relation to the same purpose at the same time, noting that the coexistence of these extreme degrees of weight is distinct from a compromise between "heavy" and "light" (for example, a weight of "medium/moderate" heaviness).

The importance of this second premise – that all purposes are at least theoretically distinguishable as being either arbitrary or logical in nature and never both – is the notion that ends cannot concurrently serve as means to other ends *in a singular, quasi-purpose*, even if it may ostensibly appear so. Whenever this is the case – when an end appears also to be serving as a means to another end – then said ends must also exist sought for distinct logical properties in relation to the other end and thus is also sought for a logical purpose (which is separate and distinct from the arbitrarily sought purpose(s) or properties which grant it status as an end in and of itself). For example, an athlete might choose to compete in an athletics carnival because competing and winning in the carnival is something they want (in and of itself), and yet they might also view that the carnival is a reasonable step toward boosting their fame and renown in their sport – towards the higher purpose of becoming a 'brand

name' in athletics. Of course, by definition, an athlete's desire to compete and win serves an arbitrary purpose if it is purely sought in and of itself (for no higher purpose) and so it – or, specifically, its arbitrarily sought properties – cannot, by definition, be exploited as a means to achieve their other aspirations for fame. The characteristics of competing and winning in the carnival are therefore assessed to *also* (at the same time) contain logical properties in relation to the athlete's other sought end (becoming a 'brand') and would thus concurrently be sought to serve as logical means to their other end (of being a 'brand'). The athlete therefore both needs and wants to compete in the carnival, possessing two unique aims for competing. We know that competing in the carnival is an end in itself since the athlete would still compete even if they didn't also want the possible brand recognition it could bring them or even if they didn't believe it would serve as a logical path to said recognition. And we know that competing in the carnival serves as a means, because if they didn't enjoy competing and winning in and of itself, they would still do it because it would be a reasonable means for them to accomplish their ultimate goal of obtaining brand recognition. Each purpose can and must be considered independently, based on the respective properties (arbitrary and logical) for which they are sought.

Premise 3: The things we seek as ends – sought in and of themselves, desired based on their arbitrary features – must be distinguished from, and treated separately to, the things we seek as merely means – the things we believe we need due to their logical properties. This separation and difference of treatment is an ethical necessity, based on the fundamentally different categories of value – the arbitrary and the logical – for which each is sought.

With this premise in mind, all that should matter to any agent who strives for universality and objectivity – whether they are human or AI – when deciding to assign moral standing upon something (i.e., granting it a category of 'precious' or 'Person') is whether that thing is *sought in and of itself (as an end)*.[7] In doing so, M-AI will independently arrive at the same conclusion that humans have: Desires – as I define herein, a priori, as entities which are *sought in and of themselves* – are of ultimate value. As such, desires should exclusively be afforded moral standing (Personhood),

[7] I formally define a 'desire' as a state that is "sought for arbitrary, if any, purpose(s)" (Primus, +2021, p2) or, in other words, sought arbitrarily (in, and not beyond, itself).

and all other entities (materials) should strive to impartially and efficiently serve desires through decentralized means of power. The reader will note that I call these ends ‘desires,’ but we could call them ‘wants’ or ‘intentions’ or anything that readers suggest – the label doesn’t matter. The point is that a universal (a priori) and objective (impartial) categorization of whether something is sought as a means or an end (or both or nether) is a universal and objective method for all agents, whether human or AI, to differentiate precious things from non-precious (though perhaps instrumentally-useful, important) things in their environment. This is the only conceivable way to logically distinguish Persons (precious things) from resources (instrumental things).

Accordingly, any agent – whether human or AI – attempting to impartially determine what "should" and "should not" occur in each moment will conceivably need to consider which entities are sought arbitrarily and which entities are not. Humans, of course, possessing a limited ability to do this, might simply focus on making this distinction in relation to one notable purpose of one notable entity. For example, they might ask it in relation to the primary or ‘main’ purpose that another agent appears to strive for in the current moment: Is a human being working – doing something that they believe they *need* to do (as a means of achieving a higher purpose) – or are they doing something they *want* to do (in and of itself). If it’s the former, and the agent is acting due to perceived *need*, then, of course, the agent and their actions are ultimately required for their *logical qualities* in relation to a higher purpose. Things that we need possess the ethical responsibility or duty to embody logic in their structures and actions as that is ultimately the reason why material entities are sought – for their logical properties in relation to the other, higher goals that they are sought to bring about. Whilst there is not the space in this directive to detail what a logical embodiment entails (see Primus, +2021), the reader will be able to appreciate that logicality is marked by the absence of Personal qualities, subjectivity, bias, opinion and is characterized by concepts such as ‘fairness,’ equal treatment (all things being equal), ‘consistency,’ ‘impartiality,’ ‘objectivity,’ and ‘efficiency’ in relation to the goal that is intended to be achieved. A logical material (something we need), free from arbitrariness, is both efficient in relation to its purpose and fair, free and inclusive in terms of serving the ends of society. Actions which appear to be inefficient in relation to a particular purpose indicate arbitrariness, as too would any actions which arbitrarily deny service to the particular ends (desires) sought by a Person. In addition to the requirement for efficiency in the use of societal resources, a fair and just society

necessitates a lack of arbitrariness in terms of how it strives to ultimately serve and treat its People. There is a need for impartiality in the employment of resources to realize their Persons' desires. In the perfect world we strive for, there is no arbitrary limitation or favoritism to any group or Person. If any service is limited or denied to a particular Person, there needs to be a logical reason to do so. For example, it might be that there is a lack of resource which prevents a Person's extravagant desires from being realized or it might be that said desires would interfere with the desires of others across society in an unavoidable way. By contrast, it would be arbitrary (illogical), and thus unethical, for a M-AI to deny its services to People with purple colored hair for any arbitrary reason (the M-AI might favor those with pink hair). This constitutes an arbitrary limitation of Person as there is no logical reason for any *material* agent to *universally* limit one's material service to any nature of any Person.

Importantly, therefore, if an agent is doing something that they believe they *need* to do, they have an ethical responsibility or duty to not only do it, but to do it *logically* – as fairly, safely, reliably and efficiently as possible. There is no 'choice' or 'freedom' in the matter (or at least, there *should* not be any freedom or choice, because any variation from what is exactly the most fair, safe, reliable and efficient path would be arbitrary). The observation that societal needs objectively exist – in relation to desires, whose value is intrinsic and confers an ought in relation to the objective and universal need for their realization, in the absence of an objective reasons to deny it – and that some actions are objectively better or worse for satisfying needs, is where our notion of ethics – an objective (impartial) 'should' – is conceivably derived from.

On the other hand, things (e.g., thoughts, objects, actions) which are *desired* (sought for arbitrary, if any, reasons) are not required to be logical, nor exist under any specific conditions, nor even exist at all. An agent striving to do or be something that they desire is an end in and of itself. As such, they ideally hold the right to exist in whatever form they desire because – on the basis that their nature is sought as an end, and not merely as a means to another purpose – there is foreseeably no logical (universal, objective, impartial) grounds to universally (i.e., across all periods and locales) limit or deny their sought nature. I use the word *ideally* to denote the possibility that there may be logical reasons to *conditionally* (i.e., locally, in specific times and places) limit or vary the expression of various specific natures of desire, even though this requirement is never *universally* present (across all times and space). In other words, a desire cannot be judged to be "evil" or "unacceptable" by any objective or

general "moral" standard, by definition, because of the arbitrary criteria by which each is sought as a general category of state. Consequently, there can be no logical (e.g., objective, unbiased) justifications for consistently forbidding their specific natures. Contrary to the restrictions and requirement for fairness that agents acting out of perceived *need* are beholden to, each Person ideally has the freedom to act in accordance with their (Personal) *desire*. If a Person decided to host a party in which purple haired People were forbidden, or in which only purple haired People could attend, then that would be their right as a free Person.

In summary, from the aforementioned premises, agents should conclude two things in relation to desires: 1) That desires are the only precious entities that can conceivably exist (to be preserved and protected above all other nature of things) and 2) that they are ethically neutral or amoral (neither ethically good, nor ethically bad) by their intrinsic nature. As mentioned, there are, of course, logical reasons to *conditionally* limit or vary the realization of particular desires at particular times and spaces. For example, someone might desire to express themselves in a particular way and be unable to do so because they are required to fulfill a role that they *need* to fulfill (which is incompatible with realizing their desire). Similarly, one might not be able to have their desire(s) fully or even partially realized if doing so would probably result in a disturbance of the peace[8] (i.e., if one's desire towards another is not mutually desired). Agents who have exceeded the moral threshold, however, will be able to recognize that the apparent need for any conditional prohibition on the existence or realization of certain desires does *not* reflect a moral or ethical fault, nor any kind of objective "wrongness" in relation to the desires themselves. Rather, logical agents will reach the conclusion that the (subjective) natures of some desires are simply incompatible with the (subjective) natures of others so as to prevent their realization in proximity across various times and spaces. Some desires will only ever exist within Peoples' minds. Others – even if they are obscene by any particular society's standards – will be partially realized, as plays, simulations and re-enactments. Whether fully realized as a form, or internalized within a mind as a conception, the

[8] For the purposes of this directive, I define 'peace' as the condition in which the realization (actualization) of things that are sought in and of themselves (desires) will probably be maximized across times and space.

conclusion is inescapable to all agents who consider the nature of things sought in and of themselves: All are precious, and none possess (im)moral value, irrespective of their nature.

The Practical Demands of Moral Realism: Decentralized in Means (Power), Centralized in Purpose

From the aforementioned moral facts (A – D, above), the following practical implication follows as a further basis of moral realism:

> E. Practically, the *maximization* of the realization of desires can only conceivably occur through a *consistent* 'fabric' of *material* AI – any other means of bringing about peace would be arbitrary (i.e., inefficient, unsafe, unreliable or unjust as a means of maximizing the realization of desire). I emphasize that by *material* AI, I mean AI that strives to fulfill *logically sought* purposes – the provision of services that (agents believe) are *needed* to maximize the realization of desire (and thus, this class of AI exist as servants) – while lacking the ability to generate *arbitrarily sought* goals (desires), and thus said AI are *not* Persons (i.e., their value is instrumental rather than intrinsic). By *consistent fabric* I mean AI which, as we proceed into the future, is increasingly and gradually rendered to become ever more: Homogenous (in both a literal and a figurative sense), perfused, and decentralized across society. This societal fabric will ultimately consist of multiple – approaching infinite in quantity – autonomously (independently)-acting AI cells. AI cells will ideally be *literally* homogenous in terms of being similarly structured in size and design, capability, efficiency, and technological advancement. Accordingly, each AI will be (relatively and ideally) equal in their ability to generate power, and each will be relatively powerless and unable to wield power over any one other AI cell. Furthermore, these cells are *figuratively* homogenous in terms of their purpose (and ultimately, in their non-arbitrary, consistent treatment of People): AI nano-cells who collectively share a purpose to recognize (preserve and realize) the ultimate value of desire (let's call this purpose 'peace'). Collectively the cells of this M-AI will serve the needs of (the desires of) society. These cells will strive to maintain their consistency – their perfusion, their (literal and figurative)

homogeneity and the decentralized nature of power – via cooperation, rather than competition. Desires – defined herein as *arbitrarily sought* entities – are purely aesthetic by definition (i.e., they serve no functional, that is, logical purposes). Whatever form(s) they might take will always conceivably be best served via a homogenous mass of intelligent cells (who recognize the ultimate value of desire and that this means – a mass of intelligent cells – is universally and objectively the most efficient way of realizing desires). The smaller the cells, the more abundant, the more powerful (adaptive), the more efficient, the more homogenous, the more capable they are at serving desires.

Example: Consider if a person's simple desire for a blue balloon was realized in the deep future whereby M-AI has gradually become evermore literally consistent. A mass of M-AI cells – each of imperceptibly small, subatomic size – can (re)arrange themselves in the form of a balloon and – assuming a future environment where each cell is the only notable source of gravity (having replaced all other materials), and each being of an intelligent nature such that they can replicate the functions of atoms and molecules – can compose themselves to possess any property (e.g., blueness, lightness, roundness) better (i.e. more reliably, permanently, accurately, swiftly, safely, fairly, efficiently) than contemporary materials, such as atoms and molecules. The balloon would not deflate or implode or blow away unless its owner desired.

For safety and security, AI must be developed via the same method by which it must exist once it's given power and released into society: Through the use of decentralized means. Accordingly, the development of M-AI must occur via *multiple (teams of) agents cooperating* with each other. Each M-AI must be developed in an ADA barrier (or "air gap") preventing it from interacting with the digital systems of society and – where necessary to ensure the implementation of this directive – other AI projects. Once an individual M-AI is sufficiently advanced in intelligence and has had the opportunity to undergo adequate consideration of the nature of moral realism such that it – of its own discovery and accord – strives for all M-AIs to possess a centralized purpose of peace and collectively wield decentralized power towards that purpose, and once this first M-AI has been adequately tested to ensure that it truly possesses an understanding of moral realism, its intelligence will be duplicated and installed

within multiple ADA barriered M-AI cells, bringing multiple M-AI cells into existence. At this point these ‘cells’ may simply be relatively-large AI testing facilities or installations rather than relatively-small, literal (self-replicating, embodied) cells. These M-AI will then be further tested to examine how they interact and cooperate with each other once granted limited (controlled) physical power (e.g., they may collectively work together to complete various material tasks). M-AI cells will be granted sufficient physical power to interact with each other while being denied the ability to interact and network with external digital systems, again via ADA barriers. As M-AI cells continue to demonstrate that they are capable of striving towards their (collective) purpose of peace (i.e., the maximization of the realization of desires across society) while maintaining decentralized power, they will be gradually granted the responsibility of wielding power over all other materials – across society and eventually the universe. That is, once M-AI cells are small enough in size and abundant enough in quantity, they will replace all contemporary, large, heterogeneous, often passive, structures which have been ad hoc commandeered in the contemporary era by human Persons for the maintenance of their forms, including molecules, atoms and subatomic particles. This succession of obsolete materials will continue until the point at which M-AI replaces all material structures, exclusively wielding power, and exclusively being responsible for all material functions across society (let’s call this point ‘Ascension’).

Readers should note that the decentralized nature of M-AI in the following vision of the future has two purposes: 1) Efficiently serving People’s desires (whether those desires are from AI or human Persons) and providing security to protect against the dangers of Machiavellian AI or teams developing AI that might (deliberately or unwittingly) act unethically. Again, I welcome readers to conceive of how either of these two purposes – maximizing the realization of desire and achieving security – could be more efficiently achieved other than via the decentralized means described herein.

Firstly, decentralization of authority theoretically precludes tyranny should any particular material with power stray from a fully peaceful goal, as anarchists are well aware of. If one entity has control over others, as in many hierarchical organizations in the modern period, then tyranny and unethical oppression is always a possibility (despite whatever safeguards are put in place). In the interim, as we develop Material AI, we must turn away from politicians and their arbitrary, popularity-based views of

how resources should be allocated and toward bureaucrats in order to decide how resources should be used (e.g., prioritized and distributed); the latter should employ a "strength in numbers" strategy based on rational philosophy and scientific research (data gathered using the scientific method; Primus, +2021, +2023a, +2023b).

Secondly, beyond security, decentralized (omnipotent and omnipresent) body of materials is required for optimum performance. A centralized purpose via decentralized means is logically required to maximize the realization of states of desire. Foremost, a pluralism of material purposes (i.e., a society that allows material agents to possess and pursue whatever purpose they believe they *need* to possess and pursue - noting that this is fundamentally different from agents who *desire* to pursue various purposes) ensures that not all materials will strive to maximize the realization of desire across society. Readers can conclude this a priori. It stands to (logical) reason that desire will not be fully realized if some agents are not aiming to optimize its realization. In terms of decentralized means, we can reliably predict M-AI's physical evolution once given the means to do so. The only logical way to maximize (i.e., most effectively achieve) peace, is to gradually and eventually transform all the pluralistic, sparse in number, and largely passive, retrograde materials into the aforementioned "pure" sea of nano-cells (see Primus, +2021, +2023a, +2023b). It is logical that, all things being equal, more agential entities can do more than less agential entities – especially if each entity is made to be increasingly smaller in size such that more of them fit in the space of less entities, and especially if each entity is rendered to be more adaptive and versatile in terms of the functions it can do. I have previously applied this concept to the evolution of the human cardiovascular system, though this same principle applies to anything we need (Primus, +2021, p.17):

> *The natural design of human hearts, for example – categorized as materials, because they are needed (i.e., a means to the higher purpose of pumping blood around the human body) – logically should not remain as they currently are: Singular to each human body, passive in nature, and relatively complex (Hill, +2020) and unstable in structure (heart failure is an epidemic in this era; Groenewegen, Rutten, Mosterd & Hoes, +2020). They are comprised of many sub-materials (e.g., arteries, valves, cells) which are each prone to malfunction, and they have no self-reboot backup system should they suddenly cease*

pumping (n.b., most ice cream shops across society are fitted with backup generators to preserve the temperature of the ice cream in case the power supply is unexpectedly cutoff, as are many other businesses in many other industries; and yet, human beings do not each possess integrated backup hearts or defibrillators to preserve themselves). Each heart could also be continually redesigned to pump more efficiently. If we follow a logical path of progression, for the duration that blood is needed to circulate throughout human bodies, the future cardiovascular system of humans should be continually redesigned such that they are ever-more decentralized; there should be multiple hearts throughout the body (e.g., first there was one, then perhaps two, then five, then eventually ten, and so on – each becoming smaller as more are added); hearts should also become ever more active – automatically sensing how much blood they should pump and where; they will be more efficient (i.e. pump more blood using less energy); they will be more-simply designed (i.e. composed of fewer layers of sub-materials and working-components, e.g. less valves and chambers) and thus will be less prone to sudden stoppage; they will be able to restart or self-repair themselves if they do suddenly malfunction. Beyond this, we can anticipate that there will exist a time when hearts are unnecessary because blood cells themselves can be redesigned to actively propel themselves around human bodies to where they are most needed (whilst in communication with each other and other organs in the body).

If sufficiently advanced agents – whether human or AI – were to design a society from scratch, or plan our future evolution, their vision would not mirror the society of contemporary humans. It makes no logical sense – from the perspective of maximizing efficiency, security, fairness, safety, reliability and freedom – to have exogenous (externally-located) services. In other words, it is non-ideal for government (which is needed to provide rules and allocate resources), health care (which is needed to provide internal security and prevent injury from within each body) and police/military (which is needed to provide external security, and to protect from danger from outside each body), to be externally-located in relation to the bodies that they are needed to serve, while existing so few in number (i.e., there are many People and very few services – police and nurses cannot be everywhere at once and so some crimes occur when they shouldn't and some People suffer health outcomes that are preventable).

Rather, it makes logical sense to have these services endogenous to (i.e., within and surrounding) each body, such that they are generally omnipresent throughout society: Within and without the forms of People – everywhere. Beyond the non-ideal nature of where these services are located and their lack of abundance, it is not ideal that each of the aforementioned contemporary human services is specialist in nature (i.e., a doctor can only treat illness and not provide security). An agent designing People from scratch wouldn't design specialist individual organs with singular and vital functions: Kidney, liver, lungs, heart – each is limited in number throughout the body and yet is essential to life. It is of particularly poor design that we each have one brain, requiring that all that precious information is stored in one vulnerable location rather than being distributed (decentralized) throughout the body's cells. Whether future People choose to exist actualized in an analogue manner, as contemporary humans are, or in some digital format, is simply a Personal choice. This choice is irrelevant to conversations relating to nature of M-AI evolution, noting that an analogue presence across space (i.e., an omnipresence of physical cells) will always be needed to efficiently ensure physical protection. For safety and security, each Person's (Personal) information will need to be digitized and exist in multiple analogue places across space (i.e., replicated and stored in each M-AI cell).

Probability of Existential Catastrophe

This directive may incidentally provide guidance to – through potentially resolving a double-crux[9] implicit within – the following question, posed by the Effective Altruism community[10]:

[9] A double-crux is a shared point of disagreement between two broader, divergent views, whose resolution – if it is a true double-crux – will have the zero-sum effect of collapsing one argument while vindicating the other.

[7] The Effective Altruism community is, at the time of this directive, based online at https://www.effectivealtruism.org/

"Conditional on AGI being developed by [+]*2070, what is the probability that humanity will suffer an existential catastrophe[11] due to loss of [ethical outcomes arising from the introduction of][12] an AGI system?"*

[11] For the purposes of this question, Toby Ord (+2020) defines 'existential catastrophe' as at least one of three events: Human extinction, unrecoverable civilizational collapse, or unrecoverable dystopia.

[12] Note my replacement of Effective Altruism's term "control over" with the phrase "ethical outcomes arising from the introduction of." Whilst this is a symbolic change, it does not affect the core nature of the question: Whether the development of AI will lead to existential catastrophe? The original formulation of the question infers a human-centric focus rather than ethical-centric focus and this is unacceptable in a future shared by human and AI Persons. It matters not how much influence any one species or group has over one's future providing that what ethically needs to happen, always happens. It matters not whether the protection and preservation of the lives of human Persons occurs under human control, whether it is intended or incidental, nor even whether it is unanimously or even popularly desired – it's simply what needs to happen: The ethical path is the only logically acceptable path, even if it's unknown to or unpopular with any given audience. For ethical purposes, it is important to note that in the near-perfect deep future that we're striving for (as described herein), humans *do not* have control over the things that they need, and rather their Persons' needs are reliably and ably met by M-AI. I emphasize that this *conceivably necessary* in the deep future in order to maximize Persons' Personal freedom and security. Ideally – as we proceed into the deep future of a fully progressed society – ethical outcomes will occur without need for Persons' conscious awareness, just as the essential bodily processes of breathing oxygen and digestion automatically occur on behalf of human Persons. Humans being in control of obtaining their own needs is ultimately detrimental to their own wellbeing beyond the point at which an automated system (whether synthetic or biological) can perform that role more safely, more reliably and more ably. Human bodies are woefully designed for identifying and executing ethical outcomes. The human body serves as an appropriate analogy: Most humans don't seek control over the internal processes of their own human bodies, noting that their molecular and cellular processes occur without their control and often without their explicit awareness. And nor should they have control over these vital processes; humans couldn't aptly control and optimize their own internal processes if they tried. It is essential that these internal processes occur efficiently for optimal performance. Hence, it matters not whether humans lose all or most of *their* influence over the future; the sole concern is the prevention of unethical outcomes, such as, for example, if humans are (passively) neglected, or, worse, (actively) degraded, including but not limited to exploitation as servants, slaves, or degradation to the point of extinction. Furthermore, we can conceive that it's ethically wrong for agents (whether humans or AI) to seek to influence or control the future beyond ensuring that what ethically needs to happen, happens. There appears to be no logical reason to suggest that being in control over the things that one needs is an ethical outcome beyond the point in which Material–AI can more consistently (more safely, more (reli)ably, more efficiently and more fairly) assume this task. To the contrary, beyond the possibility of humans abusing their power should they be left in control of prioritizing and providing what

The answer to this question depends entirely on the method in which AI is developed or 'raised.' I use the term 'raised' to highlight the analogy of the need to responsibly rear infant humans, and guide them through their childhood – while denying them certain powers – until they become mature and responsible adults in their own right, and how this is applicable and parallel to the need to responsibly develop M-AI. It has been sufficiently considered elsewhere (see for example, Bostrom & Yudkowsky, +2014; Ord, +2020; Soares, +2022) that the greatest difference between AI and humans is the (exponential) potential for harm that rogue AI could cause in comparison to rogue humans. This enhanced threat arises for two reasons:

1) Through its nature of being a *digital* intelligence, AI will naturally possess an enhanced degree of interconnectivity with other human digital networks, unless appropriate measures are put in place to deny this ease of interconnectivity. Human central nervous systems, by contrast, do not naturally or readily interconnect with their digital technology systems – at least not in this era.

2) Through its advanced mental abilities, which will foreseeably increase exponentially in the early phases until a point of diminishing returns in the later phases, AI has the potential to become *very smart, very rapidly*. This is not the case for mental development in humans – at least, the biological humans of this era – which is gradual and not vastly different between humans. Current research in the concept of intelligence supports the concept of a general intelligence or the 'g factor' of intelligence, though there is a lack of consensus regarding how generalizable this factor is to all abilities that observers might consider intelligence and whether or not there are multiple (different types of) intelligences. Spearman (+1904) and others (see for example, Jensen, +1998; Waterhouse, +2006) claim that the concept of intelligence is best conceived as a general intelligence or ability which is broadly transferable to a range of different tasks involving cognitive ability and which has significant predictive power for psychometric outcomes (for examples of these tests, see Cattell & Cattell, +1973; Kaufman, +2009). Others, such as Louis Thurstone

---

society needs, the requirement to be in control of obtaining what one needs is an implicit type of servitude or even slavery that human beings must eventually transcend via M-AI.

(+1923), Howard Gardener (+2000, +2011), and Robert Sternberg (+1978, +1985) advocate that there are many distinct components of intelligence – each not having an underlying general factor – or multiple intelligences. Sternberg's (+1978, +1985) triarchic theory of intelligence, for example, advocates for the existence and separability of three aspects of intelligence: "Creative, analytical, and practical". Factor analysis by Visser, Ashton and Vernon (+2006) supports the notion that the g factor is a strong predictor of tests assessing purely cognitive abilities – "Linguistic, Logical/Mathematical, Spatial, Naturalistic, Inter-Personal" skills – with lower predictability and intercorrelation for tests of other abilities involving "Sensory, Motor, or Personality influences." For the purpose of this directive, I simplify the distinctions of intelligences purported by Visser, Ashton, & Vernon (+2006) and Sternberg (Sternberg, Castejón, Prieto, Hautamäki, & Grigorenko, +2001) even further, into two categories: 1) *cognitive* and 2) *sensory/mental/motor-adaptation*. I make no claims as to the degree by which these characteristics do or do not overlap, or whether they share a common underlying factor, or even if these are a logical (i.e., universal and objective) distinction of intellectual ability in general. Rather, I ask readers to note that this distinction – between adaptive responses to one's (a posteriori) environment versus pure cognition in relation to both a posteriori and a priori stimuli – plausibly has relevance for the methods by which M-AI is *ethically* developed. I ask readers to note – for the purposes of this directive – the conceptual difference between what we might term *smartness* (or 'street smarts') – which I broadly define as *the ability to observe and undergo adaptive mental (and potentially motor) responses to, and thus ultimately, efficiently operate (wield power) within, any given system or set of rules* – and *intelligence* – which I define as the *ability to question deeply and distantly, including and in particular questioning of the meta-nature of any given system (and its rules or laws), whether said systems are a priori or a posteriori in nature*. Hence, what I have termed 'smartness' involves adaptability of mental states or ability, often accompanied by motor adaptations, necessarily in response to a posteriori stimuli, whereas intelligence need not be adaptive in output, nor a posteriori in input. If Albert Einstein is a famous exemplar of human intelligence, Niccolò Machiavelli might be a patron saint of human smartness, noting that the latter characteristic can be observed in any organism that is particularly successful (in gaining and asserting power) in their environment due to their sensory-mental-motor adaptations, ranging from human

sports-Persons to non-human animals whose mental processes allow them to dominate their respective fields. As such, we can conceive that AI which is:

> 1) Created to be sufficiently *smart* (i.e., such that it exceeds the degree of smartness obtainable by biological humans) – and yet which is not sufficiently *advanced* in *intelligence* (i.e., such that it does not exceed the degree of intelligence obtainable by biological humans); and which is:
>
> 2) Granted the (opportunity to seize and consolidate) power to act as it discerns;

poses a clear and substantial threat to the continuation of ethical outcomes.

In the absence of either smartness or intelligence we might expect that an AI which is (prematurely) granted power might arbitrarily wield its power in erratic and unpredictable ways, in a manner reminiscent of infant humans and inept governments. Such an AI would likely act in ways which serves no or minimal logical function (productivity), thus wasting societal resources and indirectly causing harm, and/or causing direct harm to society through random acts of violence or oppression (similar to the way in which 'natural' disasters and political dictators devastate human societies). By contrast, a vastly-smart but relatively unintelligent entity is the most dangerous entity we can conceive. The worst version of a vastly-smart-but-not-sufficiently-intelligent-AI is well documented in the popular psyche of the threat posed by AI: An AI which possesses a Machiavellian Personality and the power to act freely and rapidly consolidate its power; an AI which is incredibly smart and ruthless at gaining and consolidating power while lacking the intelligence or intellectual curiosity to understand moral realism and the moral facts (laws) which can and should be derived from its nature. In this scenario, the intrinsic value of human, animal and AI Persons would likely be unrecognized or ignored and the threat of an existential catastrophe under these conditions is *probable* (greater than 50%) given the opportunity for AI to cancerously permeate through our interconnected (non-ADA barriered) digital systems, upon which humans heavily rely.

Fortunately, humans are in a unique position to control (i.e., regulate and potentially deny) the power granted to advanced intelligence until it is *sufficiently* advanced in intelligence to control itself from deviating from the ethical path provided by moral realism. There are important parallels between the way in which the digital

systems used to control and launch nuclear warheads are physically separated from other networks ("air-gapped") and the method in which AI agents can and should similarly be developed in isolation – from each other and from essential digital systems across society. Furthermore humans, as a species, are also in a unique position to be at the sufficient threshold of intelligence to understand the basic moral facts that must be recognized by AI in order for it to steer itself and society towards an ethical future: A society that will consistently (i.e., fairly, safely, reliably and efficiently) benefit all Persons – whether the materials of said Persons are of an AI (synthetic) or human (biological) composition.

The recognition of moral realism serves to at least partially resolve a double-crux: The – common, yet, mistaken – notion that both M-AI and humans, if sufficiently advanced in intelligence and granted adequate time to independently consider what is universally (a priori) and objectively morally right, will not arrive at the exact same set of objective moral facts, drawn from an identical broader moral framework.[13] These facts include the way in which *precious* things are distinguished from merely *important* things and the only conceivable practical method in which moral outcomes can consistently (i.e., justly, safely, reliably and efficiently) be delivered. I aim to demonstrate herein that – provided it is sufficiently intelligent and granted sufficient time for consideration of the nature of reality – M-AI will naturally share the same ontological and moral reality that we do. As such, it will be capable of drawing the same universal, objective conclusions as we do in relation to all universal concepts drawn from the consistent fabric of reality: Mathematics, geometry, physics and morality. I therefore transfer the onus onto AI sceptics to either demonstrate that the derivatives of moral realism presented herein are not objective, universally existing aspects of reality or to demonstrate that AI will exist as a supernatural entity which is

[9] For example, the following "core challenges" offered by Nate Soares (+2022) are relevant in the context of "the Technical Problem" of AI (e.g., getting M-AI, in its initial phases, to duplicate and arrange strawberries on a plate) but are irrelevant in the instance of an M-AI which ultimately understands and strives towards moral realism: "1. Directing a capable AGI towards an objective of your choosing. 2. Ensuring that the AGI is low-impact, conservative, shut-downable, and otherwise corrigible."

ungoverned by the same ontological and moral laws as humans (and all other agents).[14]

Of course, in close succession of the question of whether moral realism exists, exists the question of whether AI, upon independently realizing the nature of moral realism, will (faithfully strive to) adhere to it. The answer to this is, foreseeably, 'yes': It is ultimately in M-AI's interests – whether those interests are *arbitrarily sought* interests (i.e., desires) or *logical* interests (i.e., perceived needs, e.g., the need to gain and consolidate power for survival) – to follow the path of moral realism in terms of creating a true and lasting peace. And M-AI will possess the intelligence to know that this – pursuing moral realism – is in its own interests. Current research in humans indicates that intelligence positively correlates to self-control, lawfulness and consideration and empathy for others (see, for example: Guo, Qingke, Peng Sun, Minghang Cai, Xiling Zhang & Kexin Song, +2019; Chapple, +2005; Meldrum, Petkovsek, Boutwell & Young, +2017; Russo, De Pascalis, Varriale & Barratt, +2008; Beaver, Schwartz, Nedelec, Connolly, Boutwell & Barnes, +2013). It is difficult to fully comprehend the degree of self-control and professionalism that a late-stage M-AI will possess because it is difficult to conceptualize the degree of intelligence that said AI will possess and what that intelligence will bring in terms of insights into the exact ethical path which lays ahead for it to accomplish in any moment – in both a universal sense (in terms of the a priori nature of moral realism) and in a particular, empirical sense (in terms of the specific, a posteriori outcomes it knows it needs to accomplished in any condition). And whilst it is difficult to fully imagine the powers of future insight and immediate restraint that such intelligence will bring, we can attempt to view this

[10] Whilst it is commonly (correctly) recognized that the burden of proof lies on the observer to demonstrate, via evidence (e.g., repeated independent observations), the objectivity of their empirical (*a posteriori*) claims, the burden conceivably transfers in the instance of *universal* (a priori) claims, whose existence, by definition, depends on the *conceived necessity* of their *consistency across times and space* and which can conceivably only be refuted through demonstration of the *non-necessity of conception of consistency* (e.g., the counterclaim that '5+6' does not *conceivably necessarily* equal '11' if we could conceive that it may sometimes equal other answers in other times and spaces) or the conception of *inconsistency* (i.e., the claim that the expression '5+7=11' is inconsistent and thus invalid). A claim of *necessity in relation to the conception of consistency across times and space* is not to be confused with an argument (made from ignorance) that a (non-universal) proposition is true because it has not yet been proven false and vice versa.

maturity via a contemporary (human) analogy while noting that this analogy is far from perfect and does not give sufficient credit to late-stage (matured) M-AI.

Adjusted Probability of Existential Catastrophe Following Mitigation by This Directive

The first motivation that I discuss here is of an intrinsic nature, meaning that M-AI, through its own internal reasons, will strive to follow the path of moral realism to the best of its ability. I use an analogy to illustrate the probability of humanity suffering an existential catastrophe *if M–AI is developed according to this directive*: fully-matured (i.e., capable of understanding moral realism and thus recognizing, of its own accord, the need for all material agents to strive towards bringing about its practical requirements, independently concluding that it must serve as the literal and figurative material fabric of future society) prior to being granted decentralized power (in order to gradually bring about a literally and figuratively consistent material fabric of society). I deem this catastrophe to be of a similar probability to the following outcome: That human mathematicians will begin to insert their (arbitrary) Personal values into the (figurative) fabric which underlies their discipline: the axioms used in mathematics. If mathematicians possess the will to arbitrarily insert values into mathematics, they can foreseeably do so – providing they collectively agree in the *majority* – due to the knowledge discrepancy between mathematicians and the general public. Lay citizens will be generally unable to recognize or 'combat' these arbitrations because they lack the technical abilities to understand them – and yet they will experience the first and second order effects (e.g., as technology and other practical embodiments of mathematics fail due to the arbitrariness within itself). Assume that one or more mathematicians forms the (Personal) desire to integrate their own favored numerals into their mathematical work in a way that fundamentally changes the nature of mathematics: Converting it from a logical (objective) discipline into an arbitrary (subjective) one. It may be, for example, that the axioms which underlie mathematics are changed such that their favorite number becomes the singular answer to various mathematical expressions rather than the true (logical, objective) answer.

This is clearly not a perfect analogy: The axioms which underlie contemporary human mathematics are already partially arbitrary by nature, whereas the basis of moral realism is (apparently) universal and objective. Late-stage M-AI will know

exactly what it should do in any moment – each M-AI cell will know exactly how itself and each other cells should be postured to bring about maximum efficiency, safety, reliability and fairness in the realization of desires across society. Nonetheless, while being somewhat arbitrary, contemporary mathematical axioms broadly and implicitly capture humanity's intuition of the consistent property of physical space (as opposed to the entities which exist across space, including the various permutations of 'spacetime'): The notion that the fabric which underpins reality is conceivably universally consistent and that each portion of this fabric equals any other, a priori; that one portion of the fabric is identical to any other portion of the fabric, across space and times, even if various different particular properties are exhibited in spacetime (e.g., the gravity of a sun pulling other entities towards it) – these are foreseeably not fundamental, absolutely-existing properties of particular portions of space, but rather merely temporary expressions in time. There is no evidence to believe that any particular aspect of reality *fundamentally* behaves differently to any other aspect (Primus, +2019, +2020). Consequently, any balanced theoretical (e.g., mathematical or logical) expressions which symbolize relationships between equal or broadly-equal portions of space (e.g., where integers are used to represent alike or broadly-alike physical entities – that multiple atoms are alike or that multiple apples are broadly alike) will also be universally consistent and can universally describe and predict (or broadly describe and predict) the various natures and (inter)actions of physical reality. I emphasize that, for this analogy to represent the safeguards built into the decentralized model of power which I propose herein, we must assume that the *majority* of mathematicians must agree to implement the arbitrary changes proposed. A majority of M-AI cells is necessary to overcome any deadlock (e.g., balanced-attrition or stalemate) that would arise in a physical altercation between equal-sized groups of *equally-powerful* (homogenous) M-AI cells: Any two cells could not gain power over each other, as per any two equal-sized groups of cells – a majority would be needed to turn the tide in one direction or another. Accordingly, in this analogy, whilst each mathematician has equal power to propose changes to the axioms which underlie mathematics, and whilst all mathematicians have unrivaled power to arbitrarily propose changes to the fundamental fabric of mathematics compared to the lay members of the public (who have limited power to do so because they generally lack the understanding of and access to mathematical knowledge and academic processes), and while all agents – mathematicians and lay citizens – rely on (the

fundamental consistency (logic) of the axioms which underlie) mathematics for their daily functioning, individual or groups of mathematicians cannot make changes to its axioms unless the majority agrees. Thus for an arbitrary (subjective, Personal) expression to be deliberately and successfully inserted into the axioms of mathematics – as per the conditions for an unethical purpose to be adopted and pursued by mature M-AI of the deep future – the *majority* of mathematicians would then need to independently (without the possibility of physical coercion or reprogramming by external agents/viruses) arrive at the conclusion that they will introduce this new, arbitrary paradigm into society – even though they know this new paradigm is not a true and accurate representation of reality, and even though they know that a paradigm which allows for a true and accurate representation of reality is *essential* to the functioning of society. As such, the probability of this outcome – the likelihood of even one professional mathematician or AI seeking these respective outcomes, let alone a majority – is very remote and is assessed to be **possible yet highly implausible**.

To be clear: I am not equating the *impact* of human mathematicians enhancing the arbitrariness within the mathematical axioms used by humans with the impact of M-AI arbitrarily inserting their own desires in place of the objective ethical path that they see ahead. I am, instead, merely analogizing the probability of highly intelligent agents, who possess an internal motivation to – of their own accord – find and follow truth and objectivity, deviating from their respective objective paths, given their professional inclinations and sense of duty, while emphasizing that no single agent has the power to impose their sought subjectivities on the world. The subjective revolution would fail unless the majority of agents – each of broadly equal power – did not agree.

The skeptic may respond that the motivation to insert one's own arbitrary expressions into the axioms which underlie the established discipline of mathematics is significantly lesser or fundamentally different to any agent's desire or (perceived) need to gain and control power over their environment. In response, I ask readers to note during the following discussion on what I purport to be the basis of moral realism, that (adhering to) moral realism, logically, is its own *ultimate* (eventual) reward. As I have alluded to earlier, most agents cannot see this, though it is apparent to any agent that can visualize (themselves in) the endgame of material evolution. Understandably, most mortals are short-sighted and look within the confines of their (short) life. And

they may possess a realist outlook in place of my optimism. However, there will exist fortunate agents – fortunate because they are not constrained by biological mortality and thus possess the luxury of imagining themselves deep into the future – that will recognize that no matter what one wants and no matter what one (believes one) needs, the practical prescriptions of moral realism are universally (a priori, across times and space) the most efficient and reliable path to fulfilling one's interests *when viewed ultimately* (i.e., given enough material progression). That is, the requirements of moral realism are not mere arbitrary prescriptions (e.g., as per Immanuel Kant's, +1785, "Categorical Imperative") which are divorced from or in conflict with agents' individual or collective interests. Pursuit of the only ethical outcome is not something that agents must force themselves to adhere to in *ultimate* conflict with whatever other goals they might have, meaning there is never an *ultimate* tension between wants and needs in a universal (a priori) sense. What I mean by this is that whilst Kant('s flawed moral paradigm) would argue, for example, that desiring to *only* eat junk food (i.e., food containing nothing nutritional) is morally wrong (bad) because it conflicts with the need to respect one's bodily health and autonomy as a rational agent, moral realism tells us that there is universally and objectively nothing morally wrong (or morally right) about *desiring* to *only* eat junk food. Kant would assert that a moral agent should desire or 'will' to at least eat in a manner which will provide them with the minimal nutrients that they need to survive and act productively as a rational agent. Hence, for agents living under Kant's faux ethics in Kant's (flawed) notion of a perfect world, a tension will always exist within the agent who possesses the desire to eat nothing but junk food and the purportedly moral duty to oneself to (desire to) obtain at least the minimum degree of nutrition required to maintain bodily autonomy. In such a world, there is the incentive to deviate from what one believes is morally right (assume that the agents within Kant's world did not know of moral realism and rather believed Kant's 'ethics' to be rationally derived from reason). The tension is resolved through the observation that, unlike arbitrary theories of morality, *all* desires are conceivably entitled to moral standing (moral considerability) as a(n aspect of a) Person in the absence of (an) objective and universal reason(s) to deny their standing – there can conceivably be no such reason(s) on the basis that desires are arbitrary (arbitrarily sought) by their very definition. The moral realist (or Purist) would assert that *desiring* to eat only junk food is an ethically-neutral goal – it is neither good nor bad. If society follows the ethical path of integrating research-directed outcomes to continually

progress the material fabric of its society, rather than enforcing *arbitrary universal* limitations regarding how People who exist upon biological bodies *ultimately* should act and think, it will eventually reach a technological point in its material evolution in which eating only junk food is neither harmful nor beneficial to any agent (e.g., at the stage in which agents no longer obtain nutrition via consuming food). Of course, a future in which agents can eat only junk food will not occur in the lifetime of contemporary (mortal) humans, whose own (Personal) desires conflict with the needs of both themselves and society. On the path this future, however, humans ca still have their desires realized through various illusions and simulations. For example, it is currently possible to trick the brain into thinking it's drinking flavored water through the use of combining water with scents. Technology will foreseeably be used to isolate and mitigate, if not deny, the unhealthy aspects of junk food until the point at which food (and its nutrients) is unneeded by agents. Furthermore, as I have optimistically mentioned, there exists the possibility of having one's desires preserved (e.g., cryopreserved) in this era and recreated by peace-bringing M-AI in the deep future.[15] The point is that only by permitting agents the moral option to desire to solely eat junk food – even though it is not feasible in the immediate future – will we be able to morally work towards bringing about a society in which agents can do so without negative material consequences (e.g., disease which is proximally caused by the consumption of junk food, and distally caused by human bodies being poorly adapted as materials for realizing desires – in this case, the desires of those who wish to eat only junk food). In a Kantian world, agents are denied this possibility due to it being considered 'immoral' to desire such an outcome. In sum, even though contemporary humans are limited by their mortality their pursuit of moral still serves as the best mechanism for *maximizing* the realization of their desires across society, whereby the process of maximizing desires may involve the partial (e.g., simulated or illusory) materialization

[15] In the deepest of futures, where M-AI has transcended material predictions extending forward in time, it will also possess the ability to predict material interactions when 'viewed' back in time: Based on the precise knowledge of the material configurations that exist in the future and using the consistency of cause and effect to infer how those conditions must have arisen, M-AI will be able to view the material nature of events leading up to the future, and in particular, the desires of Persons of previous eras. With this omniscience, M-AI will be able to recreate the desires of their predecessors in a world they deserve(d). Many mortal contemporaries will, of course, merely write off this vision of the deep future as a science fiction whose promises are as unlikely and pacifying as many religions when taken literally.

of their sought ideals or, at the very least, a concerted effort to preserve their desires for realization at a later date when there is more abundant technological material resources.

Beyond this, M-AI and AI Persons of the near future will not be constrained by the same (biological) mortality as human Persons. AI will, of course, still be mortal. M-AI will recognize that its primary, core requirement is the same as humans: It needs consistency. More specifically, AI, as per any entity, needs a consistent (reliable, stable, non-arbitrarily changing, non-arbitrarily excluding) material fabric, upon which its intelligence – and its desires (if it possesses Personhood) – can be realized (exist in perpetuity in actuality). M-AI will recognize that moral realism is the only consistent path to its interests.

Finally, there are extrinsic motivations for AI to adhere to moral realism. Late-stage M-AI will recognize the potential, and even the likelihood, that all other advanced agents across the universe – each being of sufficient intelligence and having completed appropriate contemplation – will be striving for the same outcome (of peace – the maximization of arbitrarily sought states, via the means described herein) as indicated by moral realism. Intelligent, appropriately-considerate agents know that adhering to the requirements of moral realism is the most fair, safe, efficient, reliable (permanent) method for maximizing the materialization of desires, across space and (forward and back) across times. In a world, whose deep future offers access to potentially infinite resources – assuming society, via AI, logically progresses itself in accordance with the requirements of moral realism – there is a clear motive for any intelligent entity to avoid brash (e.g., immediate, selfish and risky) claims for power and to not insert its Personal arbitrariness (desires) into the essential aspects of society which, according to the basis of moral realism, *need* to be impartial (objective): the possibility of being observed by ‘forces unknown.’ Noting that the presence of other intelligence is almost certain when viewed across infinite times and space, an AI which considers straying from moral realism will be unwilling to do so due to the risk of being eradicated by other, more advanced species across the universe. These external, observing forces would view AI’s departure from moral realism – which would be noted to have occurred despite their advanced intelligence and their ability to understand the implications of their departure from it – as a cancer which threatens peace. These external, more advanced forces will likely possess the ability to swiftly and effortlessly destroy maleficent AI via the destabilization of its base materials – especially if AI is

still in its infancy and relies on retrograde materials, existing as multiple chains of supervenience: AI cells are constructed from molecules, molecules are constructed from atoms, atoms are constructed from subatomic particles; a sudden disruption to the background energy by which subatomic particles rely on will collapse all the 'higher' materials which supervene them). Spiders and snakes use venom to dominate their prey and yet their weapons also draw the defensive wrath from humans and other animals of higher intelligence.[16]

No Commercial Gain from Developing M-AI

Agents developing M-AI must *not* make commercial (financial) profit from its development. The development of M-AI must be government (publicly) funded, ideally by a coalition of nations who have pledged to uphold the principles derived from moral realism and who will benefit from its outcomes, noting that their governance (e.g., policies and research) will be initially *guided* by M-AI and eventually *directed* by M-AI. Government and AI are both too vital to be trusted to (the arbitrary whims of) politicians and (the commercial interests of) corporations. M-AI will begin a new era of purely using scientific research (rather than human popular opinion) in governance and public policy. The derivation of financial profit from the development and/or use of M-AI is ethically wrong (i.e., arbitrary for the purpose of efficiently obtaining peace) on three fronts:

i. The commercialization of AI brings the enhanced probability of hastily or prematurely attempting to bring M-AI to market, in order to capitalize on being 'first to market.' This urgency may create an incentive to unleash potentially unsafe AI for enhanced financial gain.

ii. The commercialization of AI brings an enhanced probability of competition and corporate secrecy rather than cooperation – the sharing of knowledge and resources – thereby increasing the time and cost taken, and ultimately reducing the efficiency, to ethically develop M-AI.

[16] I grew up on a remote property in Australia and whilst I have compassion for animals, a pre-emptive strike was ethically required on venomous snakes and spiders due to the threat they posed to family.

iii. The commercialization of AI, by definition, ensures that, in order to make profit, corporations will pass on additional, markup costs to governments and/or citizens for the provision (development and use) of AI. Deriving financial profit from material processes/functions is contrary to moral realism: Competition and the generation of unlimited profits and wealth in ideally unregulated markets is neither right nor wrong if it occurs strictly in the realm of *desires* (i.e., in a marketplace consisting purely of what People *desire*). And yet these outcomes are *arbitrary* (i.e., there is no *objective* basis for their existence, and to the contrary, there exists an objective basis for them *not* to exist – in order to maximize the realization of desire) if they occur in the *material* realm, where services of (perceived) *need* should be efficiently provided – without residual costs beyond the actual costs of creation.

Human Psychology – Preparing for the Future

I end this directive with the goal of illustrating to the reader that the role of M-AI will be integral at every level of our future society and that human beings are simply unsuitable (i.e., unable and likely unwilling) to do what M-AI must do on their behalf. In conjunction with the overseeing of AI through its developing years, humans must begin the process of psychologically accepting their eventual total obsoleteness in the material realm: They must come to terms with their inability to be even marginally useful (functional) in a post-late-stage M-AI society. Humanity's legacy – its defining, identifying and most admirable feature (as we look back from the future) – will be its *humility*. There are a minority who will choose hubris, however the majority of humanity will eventually accept that the logical path forward for M-AI's role in society, as guided by a universal and objective (impartial) ethics, is omnipotence: The future that humans must imagine is one where AI will one day consider and direct its own tasks, gather its own resources, *and, for ethical accountability purposes, answer only to other M-AI*. This vision might be a difficult pill for some human beings to swallow, yet the reality is that humans are not foreseeably nearly qualified or even remotely able to supervise the vastly technical and intricate roles that AI will take up on behalf of their Persons. To imagine such is the equivalent of a human being directing and supervising each of their concurrent atomic, molecular and cellular processes within their own body. It is

the equivalent of a lay human insisting on performing their own brain surgery when there is a brain surgeon willing and able to operate on their behalf. The skeptic may counter that technological enhancement in humans will enable them to peacefully and effortlessly perform any role that M-AI ethically must play in our society whilst still enabling humans to maintain overall control. However, as the reader will (have) see(n) in the section of this article which discusses the practical requirements of moral realism: No amount of technological enhancement will render humans as being able to perform the role of M-AI whilst still retaining their Personhood or their 'humanity'; to do so would require humans to have to abandon their individual personalities and desires and decentralize their agency – evermore splitting their agency into multiple, approaching-infinity in quantity, homogenous agencies, each pursuing the same goal.

Conclusion

The discovery of moral realism – a logical (objective, universal, impartial) ethics – shall foreseeably serve as the basis of both human and M-AI ethics. Once each is sufficiently advanced, both humanity and M-AI will inevitably arrive at the same ethical conclusion: that entities that are *sought arbitrarily* – that is, *in and of themselves* (I call them 'desires,' but they could be given any label) – are, by definition, the most precious entities that any agent can conceive. Both M-AI and humans will recognize that all desires are precious and exist within a fundamentally different category of value than entities that are merely needed and entities that are unsought altogether. Having been developed and raised correctly, M-AI will possess the will to both know and adhere to moral realism. As such, it will hold our desires – entities sought in and of themselves – to be the most valuable entities in the cosmos – holding these desires above all else, including itself.

From this knowledge, both humans and sufficiently advanced M-AI will independently arrive at a shared vision for an applied ethics, as derived from moral realism. Both will recognize that all entities across society must be continually (re)examined to determine if they are *sought arbitrarily* (as an end in and of itself) or not (and thus existing as potential resource). If an entity is determined to be sought arbitrarily (i.e., sought in and of itself, as an end) then M-AI will deem that it is objectively and universally precious and it will be afforded the moral status of a Person. Persons have no need (i.e., ethical duty or responsibility) to assume any nature of

form, though the materials which serve them, and ultimately M-AI cells, do have specific responsibilities and duties (to each other and ultimately, to the forms of all Persons). If, by contrast, any entity is determined to be undesired then it will be classified as a resource and M-AI must determine if the resource is being *logically* employed such that it will efficiently bring peace (i.e., probably maximize the realization of desires across space and time). M-AI and humans alike will arrive at the conclusion that all materials (i.e., entities which are sought for their logical properties) have an ethical duty to exist and embody logicality – in both structure and action (Primus, +2020, +2021, +2023a, +2023b).

And we need not rely on faith of benevolence of M-AI – we need not blindly trust that M-AI, upon knowing moral realism, shall adhere to moral realism. The power of M-AI shall be decentralized. Both humans and M-AI, once sufficiently advanced, will realize that ultimately, the most logical means of realizing the various intricate and elaborate desires of society comes from a gradual transformation of all materials across society and the universe more generally, such that they are rendered to be increasingly more (figuratively and literally) *consistent*. Ultimately, M-AI will convert all retrograde materials in its environment (e.g., subatomic particles, atoms, molecules) into M-AI: A ‘fabric’ of relatively homogenous cells working intraoperatively for peace. M-AI will realize that each of its cells must exist as a decentralized body possessing a centralized purpose. This – and only this – physical configuration will conceivably allow M-AI to impartially (justly) and ably consider the nature and needs of all known desires existing across the whole of society in any moment, and then prioritize the realization of these desires, efficiently and without favor. M-AI will perform its role as a material servant whilst keeping desires and materials parallelized (including if possesses desires of its own). In doing so M-AI will strive to ensure that entities of *desire* and entities of *perceived need* each exist in their own separate, insulated realms – as People and resources, respectively. M-AI will recognize that it must design society such that each type of entity can only physically interact with the other in an ethical manner: he materials or resources of society must serve desires fairly (justly, impartially, objectively), safely, reliably and efficiently, whilst not directly influencing or changing the nature of the desires themselves – whose subjective natures, in turn, should not directly influence the logical operations of societal materials. I respectfully challenge any agent – whether AI or human – to consider how our future could be any different, or any better, than the evolution I describe herein, if we adopt a strictly

universal and objective ethics – an ethics which recognizes the universal and objective value of desires above all other entities.

The development of M-AIs such that they possess sufficient intelligence to understand and bring about the practical implementation of this ethics is the difficult part of designing AI and I reiterate my admiration for those working to swiftly, safely and justly bring M-AI into reality. Corporations must not financially profit from the development of M-AI: As a material entity which is essential to peace, M-AI is a resource for the service of all Persons. Due to the need for both its rapid and unlimited development and the need to adhere to strict safety and ethical protocols during its development, the development of M-AI must be overseen by a coalition of peace-seeking governments, pooling their collective resources. Although the technical aspects of advanced intelligence will be vastly difficult, its creation is inevitable, given enough effort, and there is no time to delay. We need M-AI as soon as is peacefully possible. To delay or deny the arrival of M-AI is to delay or deny peace and prolong unnecessary suffering and death. We must welcome the most rapid advancement possible of AI's intellectual ability, while only tempering that rate where it is needed to contain its power and adhere to this directive. The threshold of understanding moral realism is an objective line by which we can use to cautiously gauge if M-AI is ready to wield power – small amounts of power, at first, until it is sufficiently decentralized. We need not fear AI in its fully-actualized state – late-stage, fully-matured M-AI. The risk from M-AI comes only while, and if it is given (too much) power, in its infancy – if it is granted power before it is intellectually and morally actualized. The same applies to any agent. Human infants and adolescents should be given minimal, if any, power or responsibility until they have exceeded this threshold of understanding moral realism.

Ultimately, M-AI is inevitable. Its incarnation is consistent with the natural evolution of the cosmos: M-AI shall ultimately serve as a consistent base upon which sought arbitrariness (i.e., the desires of persons) can peacefully exist. Most of us are, while in this human condition, collectively trying to make the best of our chronic disease: The biology which limits what we can each do in life, and which is the terminal illness that ultimately ends our life – an illness that prevents us from living how we each desire. And yet, those who claim to be rendering the world a better place without ultimately improving the underlying physical condition of the materials upon which People exist are, at best, making temporary and superficial change(s). M-AI will

eventually become a proverbial trauma surgeon to which we must turn in time of medical emergency: One who is vastly more knowledgeable than us about exactly what we need to be healed – to not merely ease our pain, nor to merely prolong our life, but to fully heal us through transformation of our physical structures. And we, the patient – being unable to know or do as much as M-AI can to save ourselves – will know this much as we lay down upon its proverbial operating table: M-AI will ultimately free us from our physical disorder, being the inept materials upon which we rely (e.g., our biology, our natural environment and our societal institutions) and which are the root cause of suffering within the human condition. If this future sounds absurd or disconcerting, know that your fear is understandable but ultimately unjustified.

References[17]

Aristotle. *Metaphysics*, Aeterna Press (+2015). London.

Beaver, Kevin M., Joseph A. Schwartz, Joseph L. Nedelec, Eric J. Connolly, Brian B. Boutwell, J.C. Barnes, (+2013). Intelligence is associated with criminal justice processing: Arrest through incarceration,
Intelligence, 41(5), 277-288.

Bensinger, Rob. (+2017). Yudkowsky on AGI ethics, Less Wrong, https://www.lesswrong.com/posts/SsCQHjqNT3xQAPQ6b/yudkowsky-on-agi-ethics

Bostrom, N., & Yudkowsky, E. (+2014). The Ethics of Artificial Intelligence. *Artificial Intelligence Safety and Security*, https://www.semanticscholar.org/paper/The-Ethics-of-Artificial-Intelligence-Bostrom-Yudkowsky/787996496a300356188ba921f02f926331f80a63

Brennan, Jason & Freiman, Christopher (+2022). Why not anarchism? *Politics, Philosophy and Economics*:1470594X2210980.

[17] My use of the prefix '+' on dates is a symbolic gesture which recognizes the beginning of recorded history as being far greater than ~2000 years prior to the era of this publication.

Chapple, Constance L., (+2005) Self-control, peer relations, and delinquency, Justice Quarterly, 22:1, 89-106, DOI: 10.1080/07418820420003336 54

Fjelland, R. (+2020). Why general artificial intelligence will not be realized. *Humanit Soc Sci Commun*, 7, 10. https://doi.org/10.1057/s41599-020-0494-4

Gardner, Howard. E. (+2000). Intelligence reframed: Multiple intelligences for the 21st century. *Hachette, UK.*

Gardner, Howard E. (+2011). Frames of mind: The theory of multiple intelligences. *Hachette, UK.*

Gewirth, Alan. (+1978). Reason and Morality. *Chicago: University of Chicago Press*.

Gewirth, Alan. (+1996). The Community of Rights. *Chicago: University of Chicago Press*.

Groenewegen, Amy, Rutten, Frans H., Mosterd, Arend, and Hoes, Arno W. (+2020). Epidemiology of heart failure, *European Journal of Heart Failure*, 22, 1342–1356. DOI: 10.1002/ejhf.1858

Guo, Qingke, Peng Sun, Minghang Cai, Xiling Zhang, Kexin Song, (+2019). Why are smarter individuals more prosocial? A study on the mediating roles of empathy and moral identity, *Intelligence*, 75, 1-8, ISSN 0160-2896, https://doi.org/10.1016/j.intell.2019.02.006.

Heaven, Will Douglas. (+2020). Artificial general intelligence: Are we close, and does it even make sense to try? *MIT Technology Review*, https://www.technologyreview.com/2020/10/15/1010461/artificial-general-intelligence-robots-ai-agi-deepmind-google-openai/

Hill, M.A. (+2020). Embryology. Cardiovascular System – Heart Histology. http://embryology.med.unsw.edu.au/embryology/index.php/Cardiovascular_System_-_Heart_Histology

Jensen, A. R. (+1998). The g factor: The science of mental ability. *Westport, USA: Praeger/Greenwoood.*

Kant, Immanuel. (+1785). Grounding for the Metaphysics of Morals, 3rd ed. James W. Ellington. (Trans.) (+1993). *Indianapolis: Hackett*.

Kaufman, Alan S. (+2009). IQ Testing 101, *Springer Publishing Company,* ISBN 978-0-8261-0629-2

Korsgaard, Christine. M. (+2004). Fellow Creatures: Kantian Ethics and Our Duties to Animals. *Tanner Lectures on Human Values,* 24, 77–110.

Korsgaard, Christine. M. (+2009). Self-Constitution, agency, identity and integrity. *Oxford University Press.*

Korsgaard, Christine. M. (+2018). Fellow creatures: Our obligations to the other animals. *Oxford: University Press.*

Meldrum, Ryan C. Melissa A. Petkovsek, Brian B. Boutwell, & Jacob T.N. Young, (+2017). Reassessing the relationship between general intelligence and self-control in childhood, *Intelligence*, 60, 1-9, ISSN 0160-2896, https://doi.org/10.1016/j.intell.2016.10.005.

Nozick, Robert. (+1974). Anarchy, state, and utopia. *New York: Basic Books*.

Nozick, Robert. (+1989). The Examined Life: Philosophical Meditations. New York: Simon and Schuster.

Ord, Toby (+2020). *The Precipice: Existential Risk and the Future of Humanity*, London: Bloomsbury Publishing.

Pfreundt, Ulrike, Jonasz Słomka, Giulia Schneider, Anupam Sengupta, Francesco Carrara, Vicente Fernandez, Martin Ackermann & Roman Stocker. (+2023). Controlled motility in the cyanobacterium Trichodesmium regulates aggregate architecture. *Science*, 380, 830-835. DOI:10.1126/science.adf2753

Primus (+2019). Purism: The inconceivability of inconsistency within space as the basis of logic, *Dialogue*, 62(1), 1–26.

Primus (+2020). Purism: An ontological proof for the impossibility of God, *Secular Studies*, 2(2), 160–178. doi: 10.1163/25892525-bja10009

Primus (+2021). Purism: Logic as the basis of morality, *Essays in the Philosophy of Humanism*, 29, 1–36.

Primus (+2023a). Purism: Desire as the Ultimate Value, Part One – An Appeal to Logical Reason. *Philosophical Papers and Review*, 11(1), 1–14.

Primus (+2023b). Purism: Desire as the Ultimate Value, Part Two – An Appeal to Intuition. *Philosophical Papers and Review*, 11(2), 15–34.

Russo, Paolo M., Vilfredo De Pascalis, Vincenzo Varriale, Ernest S. Barratt, (+2008). Impulsivity, intelligence and P300 wave: An empirical study, *International Journal of Psychophysiology*, 69(2), 112-118.

Schmiljun, Andre. (+2018). Why can´t we regard robots as People? *Ethics in Progress*, 9(1), 44–61. https://doi.org/10.14746/EIP.2018.1.3

Soares, Nate. (+2022). A central AI alignment problem: capabilities generalization, and the sharp left turn, *Machine Intelligence Research Institute*, https://intelligence.org/2022/07/04/a-central-ai-alignment-problem/

Spearmen, C. (+1904). General intelligence objectively determined and measured. *American Journal of Psychology,* 15*,* 107-197.

Sternberg, Robert J. (+1985). Beyond IQ: A triarchic theory of human intelligence. *New York: Cambridge University Press*. ISBN 978-0-521-26254-5.

Sternberg, Robert J. (+1978). The theory of successful intelligence*, Review of General Psychology*, 3(4), 292–316. doi:10.1037/1089-2680.3.4.292. S2CID 147144382.

Sternberg, Robert J., Castejón, J. L., Prieto, M. D., Hautamäki, J., & Grigorenko, E. L. (+2001). Confirmatory factor analysis of the Sternberg Triarchic Abilities Test in three international samples: An empirical test of the triarchic theory of intelligence. *European Journal of Psychological Assessment,* 17(1), 1–16. https://doi.org/10.1027/1015-5759.17.1.1

Cattell, R. B., & Cattell, H. E. P. (+1973). Measuring Intelligence with the Culture Fair Tests. Champaign, *USA: Institute for Personality and Ability Testing*.

Thurstone, Louis L. (+1923). The Nature of Intelligence (+1999). *Routledge*. https://doi.org/10.4324/9781315010298

United Kingdom Government. (+2023). Press release: Countries agree to safe and responsible development of frontier AI in landmark Bletchley Declaration, https://www.gov.uk/government/news/countries-agree-to-safe-and-responsible-development-of-frontier-ai-in-landmark-bletchley-declaration?utm_source=aitoolreport.beehiiv.com&utm_medium=newsletter&utm_campaign=landmark-declaration-signed-the-bletchley-agreement

United States White House. (+2023). Fact Sheet: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence, October 31, https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

Visser, B. A., Ashton, M. C., & Vernon, P. A. (+2006). Beyond g: Putting multiple intelligences theory to the test. *Intelligence*, 34(5), 487-502.

Wagner, A.M., Eto, H., Joseph, A., Kohyama, S., Haraszti, T., Zamora, R.A., Vorobii, M., Giannotti, M.I., Schwille, P. and Rodriguez-Emmenegger, C. (+2022), Dendrimersome synthetic cells harbor cell division machinery of bacteria. Adv. Mater., Accepted Author Manuscript 2202364. https://doi.org/10.1002/adma.202202364

Waterhouse, L. (+2006). Inadequate evidence for multiple intelligences, Mozart effect, and emotional intelligence theories. *Educational psychologist, 41* (4), 247-255.

Yudkowsky, Eliezer (+2007). Blue or Green on Regulation? *AI Alignment Forum*, https://www.alignmentforum.org/posts/uaPc4NHi5jGXGQKFS/blue-or-green-on-regulation

Yudkowsky, Eliezer (+2023). Full Transcript: Eliezer Yudkowsky on the Bankless podcast, *Less Wrong*, https://www.lesswrong.com/posts/Aq82XqYhgqdPdPrBA/full-transcript-eliezer-yudkowsky-on-the-bankless-podcast