AI & SOCIETY (2022) 37:1519–1532
https://doi.org/10.1007/s00146-021-01228-7

OPEN FORUM

Discourse analysis of academic debate of ethics for AGI

Ross Graham[1]

Received: 19 April 2021 / Accepted: 26 April 2021 / Published online: 2 June 2021

Abstract
Artificial general intelligence is a greatly anticipated technology with non-trivial existential risks, defined as machine intelligence with competence as great/greater than humans. To date, social scientists have dedicated little effort to the ethics of AGI or AGI researchers. This paper employs inductive discourse analysis of the academic literature of two intellectual groups writing on the ethics of AGI—applied and/or 'basic' scientific disciplines henceforth referred to as technicians (e.g., computer science, electrical engineering, physics), and philosophy-adjacent disciplines henceforth referred to as PADs (e.g., philosophy, theology, anthropology). These groups agree that AGI ethics is fundamentally about mitigating existential risk. They highlight our moral obligation to future generations, demonstrate the ethical importance of better understanding consciousness, and endorse a hybrid of deontological/utilitarian normative ethics. Technicians favor technocratic AGI governance, embrace the project of 'solving' moral realism, and are more deontologically inclined than PADs. PADs support a democratic approach to AGI governance, are more skeptical of deontology, consider current AGI predictions as fundamentally imprecise, and are wary of using AGI for moral fact-finding.

Keywords Existential risk · Artificial general intelligence · Superintelligence · Ethics · Discourse analysis

1 Introduction

AI, defined simply as "non-biological intelligence" (Tegmark 2017: 39), has re-emerged as a field of intense scholarly interest following the 'AI winter' of the latter twentieth century, when enthusiasm and funding for AI research temporarily dried up. Partly, this is due to breakthroughs in deep-learning and neural nets (Bostrom 2014), which bypass prior problems in machine perception by instead analyzing massive empirical datasets via increasingly fine-grained and sophisticated pattern detection algorithms. If intelligence is truly the "ability to accomplish complex goals" (Tegmark 2017: 50), AI has made great strides of late, provided that the goal in question is well specified and narrow.

Human intelligence remains more adaptable, able to derive and solve complex, partially specified problems across multiple domains using more general cognitive faculties. Artificially achieving a similarly 'general' intelligence is a long-standing AI research goal (Bostrom 2014; Kurzweil 2005; Good 1965; Turing 1950). However, this goal has branched off from the dominant paradigm of machine-learning-based AI research to form a somewhat distinct research program (Freed 2020). Implicit in generally intelligent AI is the prospect of a machine with equal-or-superior intelligence to humans themselves. This, coupled with the scalar advantage machines have over biological organisms (e.g., no need for sleep, food, near-zero marginal cost to reproduction), raises numerous ethical issues. The most canonical of these is Good's (1965) prediction of an intelligence explosion. Briefly stated, as machines surpass human intelligence, they can design and produce better machines than humans, which logically means self-improvement. Good therefore feared an exponential explosion in the rate of production of increasingly intelligent machine entities, an accelerated equivalent of the runaway cognitive gap between humans and other animals today.

AI that outstrips human intelligence is commonly referred to by three terms in the literature: artificial general intelligence (AGI), superintelligence, and high-level machine intelligence. AGI is defined as an artificial intelligence (AI) possessing equal and/or superior intelligence to humans, including "common sense and an effective ability to learn, reason, and plan to meet complex

✉ Ross Graham
rdgraham@ucsd.edu

[1] Department of Sociology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

information-processing challenges across a wide range of natural and abstract domains" (Bostrom 2014: 4). A superintelligence is "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest" (Bostrom 2014: 26), while high-level machine intelligence (HLMI) is AI "able to perform almost all tasks that are economically relevant today better than the median human" (Zhang and Dafoe 2019: 34). These definitions express different degrees of separation between human intelligence and the AI. Indeed, superintelligence is generally considered the stage subsequent to achieving either AGI or HLMI. The 'intelligence explosion' that concerned Good describes an AI's capacity to develop superintelligence upon reaching a general intelligence threshold similar to humans. Nevertheless, the ethical concerns of these three developments substantially overlap due to having a common moral concern: *AI possessing greater intelligence than humans*. For analytic clarity, this paper henceforth uses AGI as a catch-all term.

Many leading researchers and practitioners believe that AGI will be achieved by the end of this century (Müller and Bostrom 2016; Grace et al. 2017; Walsh 2018; Etzioni 2016). In anticipation, an increasingly vigorous ethical debate underway. Since many of the researchers engaged in this debate also develop the technology itself, their ethical perspectives are liable to directly influence the development of AGI. This development is ongoing—approximately 72 AGI projects are underway globally (Fitzgerald et al. 2020). Furthermore, while the project of creating AGI is distinctive from AI, stimulating a debate about the ethics of one is liable to also generate debate and deliberation over the ethics of the other, adding much-needed overall nuance and conceptual clarity (Prunkl and Whittlestone 2020). This is particularly true in a few very large private companies like OpenAI and DeepMind that are working on regular AI and AGI concurrently. AGI is a particular issue, but is related to the wider world of AI development.

Ethical debates among humans are imperfect affairs. An ethical belief rising to prominence is not merely a result of its moral superiority over competing ideas. Social forces affect groups who advocate for a particular ethical view, and thus can influence whether that view is adopted (or not). Sociologists are apt to probe the social forces that affect adoption of ethical principles. Evans (2002), for instance, dissected the bioethical debate over human gene editing. He observed decreasing influence among theologians when compared to scientists, a consequence of an epistemological shift from substantive to formal rationality. The substantive rationality of theologians meant that they debated both the *means* and *ends* of bioethical quandaries. Scientists, by contrast, considered only the *means* by which some pre-ordained end could be achieved, avoiding messy value-laden debates about whether the end was optimal. This allowed bioethics to fit better with increasingly technocratic governance in the US (Evans 2006).

A sociological analysis similar to Evans' of the AGI ethical debate is lacking from the literature. AGI will be the most powerful technology ever invented (Bostrom 2014; Russel 2019; Tegmark 2017), so acquiring this knowledge has great humanitarian importance, and provides direction and scope for future research. I am specifically interested in the academic debate, yet am aware AGI development is not solely an academic matter—many projects are housed in private technology companies, and projects backed by governments and militaries exist also (Fitzgerald et al. 2020). A portion of the ethical debate takes place in non-academic outlets like journalism, blogs, podcasts, and technology company position and policy documents. So why separate them? First, building AGI requires substantial academic expertise, and a majority of individuals working in the non-academic space have nevertheless received academic training. As such, academic training has influenced and shaped most of these individuals to some degree, whether they continue to work in academia or not. Second, the aforementioned survey of extant AGI projects (Fitzgerald et al. 2020) identifies two important distinctions in the approach of academic and corporate AGI projects. Academic projects' stated goals are usually knowledge production, while corporate projects seek to benefit humanity (partially through generating profit, but this is appropriately beneficial (Russel 2019)). This implies that, with respect to ethics, corporate projects have a more explicitly utilitarian and teleological bent when compared to academic ones. Corporate projects are also more focused on AGI safety—identifying and preventing risks specifically (Everitt et al. 2018)—than academic projects (Fitzgerald et al. 2020). This substantial component of the ethical debate is therefore underdeveloped in academia. This justifies separating the two debates, and suggests the academic debate requires more urgent attention.

According to Collins (2018), the expertise required for AGI can be split into two camps. AGI requires experts capable of building intelligent machines, drawn from technical and scientific disciplines like computer science, electrical engineering, and physics. This group is henceforth referred to as *technicians*. It also requires expertise on humans, for two reasons. Since the threshold of ethical concern is an AI possessing intelligence superior to humans, understanding and measuring human intelligence are crucial to know both how, and whether, AGI has been achieved (Collins 2018). Here, Collins' position is perhaps overwrought—a second more modest reason for this expertise is that humans are the most intelligent entities we currently know of, and therefore our most fruitful wellspring of empirical data and conceptual understanding of intelligence in the abstract. This type of expertise is underdeveloped—there is a dearth of social science in this space (Mlynar 2018; Irving and Askell

2019), while the humanities are underrepresented also (Freed 2020), with the exception of philosophy and cognitive science (whose engagement has nevertheless waned). Using Collins distinction, it appears AGI expertise pertaining to humans is highly, but not exclusively, concentrated in philosophy. This expert group is henceforth referred to as *philosophy-adjacent disciplines (PADs)*. The split between humanities and science evokes Snow's famous lecture suggesting that intellectual life was "increasingly split into two polar groups… at one pole we have literary intellectuals… at the other, scientists" (1959: 2).

1.1 Predictions

For clarity, ethics "involves systematizing, defending, and recommending concepts of right and wrong behavior" (Fieser 2021: 1). This paper mainly considers normative ethics—what moral thought or action should we take—and applied ethics—aspects specific to the case of AGI. Since AGI is a yet-to-be invented technology, ethical conjecture and analysis is done in an anticipatory vein. This means that both normative and applied arguments employ empirical examples and extrapolate trends from current machine-learning-based AI technologies, as well as the broader categories of computers and technology. Nevertheless, scholars suggest that AGI research and development is sufficiently siloed from AI writ large (Freed 2020; Brooks 2017) that any overlap or comparison does not meaningfully reduce the extent to which the AGI debate is about AGI specifically, rather than AI generally. It is appropriate to note that this position is contested in some quarters (e.g., Fjelland 2020).

I contextualize my results by briefly elaborating on what I expect to find. Just as Evans (2002) found that scientists favored consequentialist framing more than theologians, I expect technicians to adopt a consequentialist normative ethical framework more readily than the PADs. Bostrom (2014), perhaps the leading scholar in AGI today, has said AI programmers (i.e., technicians) are likely to embrace utilitarianism by dint of the probabilistic or game theoretic nature of computation. I expect PADs to prefer deontological approaches. A general survey of philosophers by Bourget and Chalmers (2014) found that deontology (31.5%) was overall more popular than utilitarian (24%) normative ethics. Given uncertainty in the robustness, precision, and generalizability of this data, this prediction is tentative.

Secondly, I expect technicians to take a more technocratic approach to AGI governance than PADs, whom I expect to be more democratic. Technicians are closer to the commercial and financial incentives of mass-use technologies than, since they can patent or claim them as property. Since AGI is already the subject of public criticism and scrutiny, they likely will find it prudent to limit the scope of lay/public input. Philosophers are generally egalitarian in their socio-political views (Bourget and Chalmers 2014), so I expect them to vest greater governance power in the lay public.

Moral realism asks whether "moral claims do purport to report facts and are true if they get those facts right" (Sayre-McCord 2015). While some philosophers propose moral objectivism is distinct from realism, arguing that "things are morally right, wrong, good, bad etc. irrespective of what anybody thinks of them" (Pölzler and Wright 2019: 1), it is generally taken that moral realism implies moral objectivism (Björnsson 2012). Accordingly, I treat them similarly, and use moral realism as a catch-all term. Thus far, human intelligence and tools have been unable to conclusively confirm the truth of moral realism. However, a sufficiently advanced intelligence, like an AGI, may be capable of doing so, provided the notion of 'solving' the puzzle is coherent in the first place. I expect to consider the existence of moral facts *plausible*. I base this prediction on the popularity of moral realist (56%) over anti-realist (28%) meta-ethics from the same philosopher's survey (Bourget and Chalmers 2014). I make no prediction with respect to technicians, and consider the prediction re: tentative.

Finally, I expect both groups to strongly emphasize pro-activity and urgency for addressing existential risk posed by AGI, irrespective of likelihood. Four surveys of AI experts found median predictions for a 50% chance of fully developing an AGI at 2040–2050 (Müller and Bostrom 2016), 2062 (Grace et al. 2017), 2061 (Walsh 2018), and sometime after 2041 (Etzioni 2016), underscoring my prediction of urgency. Many of the same experts explicitly identify the existential risk posed by AGI (e.g., Sandberg and Bostrom 2008; Ord 2020).

2 Method

I use discourse analysis of academic articles on AGI ethics. Discourse analysis is used to make consistent and reliable inferences from bodies of text. It is a staple of qualitative social scientific and medical research, used on media documents, interviews, academic articles, archive materials, and radio/television transcripts. Discourse analysis takes an inductive approach, appropriate when theoretical perspectives are either contested or lacking entirely. This approach constructs a 'middle ground' for further developing theory, and is sometimes referred to as grounded theory (see Strauss and Corbin 1997). The inductive approach seeks to either establish or flesh out general categories without being *ex ante* guided by them. AGI ethics has theoretical form, but is primarily a contested space, hence my inductive approach.

This study is concerned with the academic debate in AGI ethics. I draw methodological impetus from Graham et al. (2019), who conducted a systematic literature review of biomedicine's approach to climate change. Rather than synthesize all literature, these authors wished to systematically communicate what is readily and habitually available to medical professionals in their day-to-day. Accordingly, they limited themselves to core medical databases like PubMed to generate a "perspectival, rather than exhaustive, literature review" (133). This approach is supported by librarian science: evidence suggests that disciplinary databases are the starting point of the preponderance of scholarly searches by academics (Borrego and Anglada 2016). As such, I have selected two databases that approximate where academics from my two populations would begin searching if interested in AGI ethics. The database Engineering Village approximates the discursive world of the technicians. It contains over 20 million articles from over 4000 journals (Engineering Village 2017), including all entries from other major repositories like arXiv and IEEE Xplore. For PADs, the open access journal archive PhilPapers covers over 4000 journals to archive 2,246,838 books and articles, roughly four times as many as the next largest philosophical database—Philosophers Index (PhilPapers 2017). Both databases include literature that pertains to their representative disciplines (engineering and philosophy) without *necessarily* being written by only by disciplinarians or published in dedicated engineering or philosophy journals. The selection of 'outside voices' is modest however. As such, the 'perspectival' approach limits or excludes certain voices from analysis. Approximating the discursive worlds of my populations is necessarily imperfect, but this imperfection serves to appropriately reflect the ways in which experts practice seeking out and synthesizing knowledge.

Both databases index using methods analogous to search terms: Engineering Village uses 'controlled vocabulary', while PhilPapers uses 'categories'. Both these features use automated, computational categorization methods that are monitored and refined by an expert editor. Since academics are liable to use these tools, I favor using them for data. Both have search categories specifically attending to the subject matter. Engineering Village uses 'Ethical Aspects' and 'Philosophical Aspects' alongside 'Artificial Intelligence' and 'Artificial Intelligence (General)', whereas PhilPapers has 'Ethics of Artificial Intelligence'. In May 2019, I compared the utility of these curated categories to manual searching, to confirm the curation process was not severely deficient in some way. Specifically, I compared the contents of the curated categories with search results derived from manually typing the topic heading of these curated categories. Curated categories were inclusive of papers returned from a manual search, and more comprehensive. I then downloaded all articles matching the term ethics (first using the database-specific categories outlined above, *then also* manually entering 'ethic' into the search-bar embedded within the curated category page) in combination with four terms for AGI: 'artificial general intelligence', 'AGI', 'superintelligence', and 'high-level machine intelligence'. After removing duplicates, reprints, or unpublished work, documents were uploaded into EndNote for management and NVivo for analysis. Having read through all the literature and handwritten annotations, I deemed three articles irrelevant and discarded them, leaving a final corpus of $n=81$ articles and books, $n=38$ for technicians, $n=39$ from, and $n=4$ from both.

These 81 articles and books were then coded closely. This is an interpretive process—over time, inductive coding allow more-abstracted categories to emerge that describe the main phenomena contained in the dataset (Elo and Kyngas 2008). First, articles were tagged as having come from literature downloaded from either the technicians or database, to identify whether any significant divergence exists between these two groups on any one debate. Then, I coded the material that concerned ethics or an ethical position. Paragraphs or segments not discussing ethical content were ignored. These codes were then revisited multiple times to identify cross-cutting themes, which allowed me to reorganize and condense them into higher order categories, namely significant ethical debates regarding AGI. For instance, codes concerning the role of scientists, technologists, policymakers, legislators, bad actors, and public opinion in AGI governance distilled into a debate between democratic and technocratic approaches to governance. In another example, codes on normative ethics for AGI were essentially comparing utilitarianism with deontology. This repeated process of abstraction makes discourse analysis a powerful method for analyzing large bodies of nuanced, complex text.

3 Results

My results are split into two parts: areas of agreement and areas of discord. Areas of agreement are topics central to the ethics of AGI debate, but which both discourses frame and appraise similarly. These are subject to briefer discussion, since my 'two cultures' analytic framework offers little explanatory power if similar views are present in both groups. Areas of agreement weave literature from both discourses throughout the discussion. *All direct quotes are from my corpus*, while supplementary literature providing context is paraphrased. Areas of agreement are existential risk, our moral obligation to the future, human identity in the age of AGI, and the relationship between morality and consciousness. Areas of discord are where the two discourses diverged notably in their framing and interpretation of the same ethical issue. Where clear differences exist between the technicians and, discussion and analysis is extended. This section clearly demarcates direct quotes by technicians, respectively. Supplementary literature providing context appears in latter paragraphs, after the data itself are presented. Areas of discord include moral realism, the role of experts and laypeople, and the deontology vs. utilitarianism debate. Not all ethical debates are included in the final analysis. The seven I selected were those I considered most significant, either because of their ethical significance, the extent of written material dedicated to them, or where the contrast between technicians and was most profound. A theme was considered present if one paragraph or more in a given article advocated for the relevant position or view. Table 1 indicates number of articles containing discussed themes.

4 Areas of agreement

4.1 Existential risk

Existential risk is the paradigm for AGI ethics. Existential risk is defined as "one that threatens to cause the extinction of Earth-originating intelligent life or the permanent and drastic failure of that life to realize its potential" (Bostrom 2014: 15). AGI poses existential risk due to a possible 'intelligence explosion' upon becoming capable of designing and editing itself and other machines. This runaway intelligence gap makes AGI powerful and inscrutable—it could discard human beings as it deems them burdensome, it may destroy humanity through indifference or by accident, or it may view human beings as detrimental to ethical cosmic outcomes. Existential risks can be thought of statistically as a black swan (see Taleb 2007) or tail risk. Put simply, a black swan is a probabilistically unlikely event that has ruinous consequences, and thus must be judged more upon the scale of its impacts than its likelihood, as over a long enough time-horizon said ruinous outcome is guaranteed. Accordingly, practitioners "should assume that AGI may present serious risks to humanity's very existence, and carefully restrain our research directions accordingly" (Yampolskiy and Fox 2013: 14). It is important not to fall prey to motivated and anthropocentric reasoning when considering AGI existential risk, for as Eliezer Yudkowsky puts it: "AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else" (Yudkowsky 2008: 27).

Human society has lived with the reality of existential risk for upwards of 60 years (nuclear weapons, climate change), and apocalyptic themes are woven into many theological traditions, so the concept of species-wide annihilation is hardly foreign. There is no particular reason for the two discourses to conceive of existential risk differently—it is an oddly egalitarian topic, equally relevant to all and of the utmost regardless of your intellectual orientation. Yet, AGI also holds great promise for ethical and humanitarian well-being, perhaps even 'solving' other existential risks like climate change—leading experts believe a backbone of highly coordinated technological infrastructures is critical for this task (see Hawken 2017), a job AGI could fulfill. It is therefore "a sharper double-edge sword than any other. It constitutes at once the greatest conceivable source of existential risk and global catastrophic risk, and our most promising means of mitigating such risk" (Cortese 2014: 7). Both groups noted that AGI will be the "most important event in the history of humanity" (Totschnig 2017: 908). It poses an existential risk, yet we already live in a world full of plausible earth-ending threats. Unlike nuclear weapons or climate

Table 1 Number of articles containing discussed themes

Theme	Description	Technicians	
Existential risk is priority	Priority for AGI development is managing existential risks it may pose, over and above considering potential benefits	17	26
We have a moral obligation to future	AGI development should presume unborn humans and future civilization have substantive moral standing today	14	19
Conscious AGI is concern	Developing machine intelligence may necessarily bring about machine sentience, moral subjectivity, and machine agency	10	9
Discuss issues of humanhood	AGI's developmental interaction with human bodies, minds, and possible enhancements raises questions about our moral identity	25	28
Favor technocracy	Majority of AGI governance requires technical and scientific expertise, rapid decision-making should defer to experts	16	6
Favor democracy	Majority of AGI governance requires democratic public input, decision-making should be deliberative	8	12
Favor utilitarianism	Ethical issues in AGI development should be considered in terms of beneficial or detrimental consequences to humans	12	5
Favor deontology	Ethical issues in AGI development should be judged with respect to obligatory rules, principles or duties	13	4
Moral realist	Moral truths and/or facts objectively exist, and an AGI could plausibly identify them, possibly for the betterment of civilization	10	6

change, AGI contains constructive as well as destructive potential for the fate of humanity.

4.2 Moral obligation to the future

AGI both sharpens and multiplies our moral obligation towards future generations of humans. Recent scholarship suggests that humans should consider more deeply their impact on the unborn, owing to the increased productivity of all forms of capital over time (Cowen 2018), the greater qualitative and quantitative scope future generations have to morally affect the world (Greaves 2017), and the uninhabited swathes of space and time that technology might one day populate with human lives (Bostrom 2003a; Beckstead 2013). Despite extensive discussion, there was little to distinguish technicians from PADs on this topic. I find concern that AGI has no intrinsic terminus (i.e., it does not die), making it both "capable of playing a much 'longer' and more deceptive game than the typical human" (Danaher 2015: 8), while the "marginal cost of creating an additional artificial intelligence after you have built the first one is close to zero" (Bostrom 2003b: 762). The consensus that AGI is an existential risk also implies consensus on its significance for all future humans, since this incurs the termination of *all* potential lives and inherently weds AGI ethics to the unborn. As such, it is hard to conceive why PADs and technicians would view this issue differently. The open question is whether an AGI can replicate human consciousnesses, possibly creating a vast community of replicated minds and greatly expanding the amount of subjective human experience in the future relative to the amount of human organisms. This refracts the scale of moral significance, as "even if a small fraction of these lives were to exist in hellish circumstances, the amount of suffering would be vastly greater than that produced by all the atrocities, abuses, and natural causes in Earth's history so far" (Sotala and Gloor 2017: 389). Accordingly, "the creation of such artificial intellects will have wide-ranging consequences for almost all social, political, economic, commercial, technological, scientific, and environmental issues that humanity will confront in this century" (Bostrom 2003b: 764).

4.3 Does morality require consciousness?

While not fully understood, moral action and responsibility is intrinsically linked to consciousness, and whether an agent can suffer. Consciousness has many remaining scientific and philosophical mysteries. In combination with the equally nebulous topic of AGI, I construe the agreement found between PADs and technicians as shared ignorance or uncertainty, rather than agreement in any definitive sense: "while it might be presumed that consciousness is a prerequisite for agents making moral decisions when confronted with complex dilemmas, the exact role(s) consciousness plays has never been fully clarified" (Wallach et al. 2011: 180).

Both groups note that a conscious AGI raises ethical questions in two areas. First, if AGI is conscious, then to what extent is it responsible and accountable for its actions: "An important issue to consider is how, and if, conscious experience relates to ethics and morality, especially to A(G)I… some ethical theories deny there can be responsibility and/or accountability without consciousness" (Dameski 2018: 44–45). Given its intelligence, it is widely held that conscious AGI should bear responsibility for its actions, presumably for the same reasons that extant but relatively unintelligent conscious entities (animals, children, mentally impaired) are relieved of this responsibility. A common way of thinking about this evoked Kahneman's 'System 1' and 'System 2' (see Kahneman 2011) model of cognition and decision-making. System 1 thinking is rote, subconscious, and automatic, while system 2 kicks in for hard tasks requiring intentional reasoning. Consciousness and moral decisions are seen to operate similarly—most decisions (including moral ones) are subconscious, automated, and obvious (we trust children with these), but certain ones require intense subjective introspection and, thus, employ consciousness more extensively. Accordingly, "consciousness will be especially important for making volitional moral decisions" (Wallach et al. 2011: 189) in AGI. Foremost experts believe that consciousness and intelligence are somewhat aligned (see Koch 2019), supporting the position that conscious AGI shoulders some ethical responsibility.

Second, a conscious entity can suffer, so if AGI achieves consciousness it too can suffer and thus becomes the subject of moral considerations: "[the] AGI system may possess a type of consciousness comparable to the human type, making robot suffering a real possibility and any experiments with AGI unethical for that reason" (Yampolskiy 2015: 139). However, such an event also presumes the derivation a mechanistic, computational and/or scientific 'answer' to the question of what brings about consciousness in the first place. Accordingly, its capacity to experience suffering is possibly accompanied by a capacity to intentionally create novel or replicated conscious subjective experiences. This might allow it to recreate suffering, and have humans suffer within simulated environments: "[we require] recognition of an ethical prohibition against running ancestor-simulations because of the suffering that is inflicted upon the inhabitants of the simulation" (Bostrom 2003d: 9), or indeed to suffer itself on a scale greater than a normal person: "from an ethical point of view, the possibility for an AI to experience supersuffering takes precedence over the expected benefits that an AI will produce over mankind" (Beckers 2017: 90). Accordingly, for both AGI and humanity alike, a conscious AGI potentially increases suffering, which neither PADs nor technicians view favorably.

4.4 With AGI, what is a human?

AGI is expected to co-emerge to some extent in biological organisms, greatly complicating the nature of human identity. Cyborgization trends are a central issue in contemporary analyses of humanhood, including the role of smartphones, computers, wearables, and a burgeoning biohacking industry (see Yetisen 2018 for an excellent summary of this topic). I find questions raised over whether (a) a 'human', as currently instantiated, will be the locus of ethical concern by the time AGI is developed and (b) whether AGI will in fact be intricately woven into human bodies and/or minds. Aliman predicts a "global identity crisis: cyborgization could initiate unforeseen social transformation shaking the notion of "human being", "identity" and "self"" (2017: 193), one where our biological limitations "could be compensated… by increasing the percentage of the non-biological part or/and by enhancing the intelligence of the non-biological part e.g. by using narrow AI components" (195). Such concerns about human identity are necessarily downstream from the scaling problem discussed earlier: software scales extremely rapidly on near-zero marginal cost, while hardware also advances quickly when set on an evolutionary timescale. Biological organisms (like humans) are slow by comparison, and require more extensive maintenance in the form of food, water and rest. Accordingly, AGI will inevitably encroach upon human identity with inhuman speed, raising questions like ""At what time does someone stop to be a human?", "Does the self include the machine part?", "What happens if the non-biological part starts to prevail—does the cyborg become machine"." (Aliman 2017: 193).

These questions are somewhat obvious, and leave little room for differing interpretations between technicians and PADs. Indeed, merging human minds and/or consciousnesses may well become the only way "feasible for the person to keep up with superintelligent machines" (Yampolskiy 2013: 405)—a similar observation drives the most famous AGI prediction, Kurzweil's singularity (see Kurzweil 2005). I do find a kind of shared pragmatism between the two populations, a result of the high degree of uncertainty and unpredictability in this topic and epitomized in the following quote: "my prediction is that instead of a "war" between humans and AAs [artificial agents], most likely, humans will become eager to "upload" themselves, get cyborgized or completely mechanized, potentially immortalized, eventually reaching a level where the differentiating line between artificial and natural, AAs and humans will become totally blurred" (Galanos 2017: 587).

5 Areas of discord

5.1 The role of expertise and democracy in AGI governance

I find that technicians favor greater top–down control of AGI governance than PADs. AGI is likely to create unpredictable socio-political dynamics between different groups, companies, nation states, as well as between humans and AGI itself. Accordingly, AGI governance is a complex ethical issue detached from whether the AGI itself is 'ethical', epitomized in this philosopher's quote: "even if the machine/AGI ethics problem were 'solved', the deployment of AGI systems on a large scale may not actually lead to positive social outcomes" (Brundage 2014: 367). Humans will remain as morally flawed as ever. This issue raises two questions that the groups interpreted differently: what role should democratic or technocratic governance mechanisms play? What is the socio-political role of the expert and the layperson?

I find technicians emphasize limitations to democratic deliberation for AGI governance, as "the sort of cautious, patient measures needed to deal [with AGI]… are alien to our [human] nature" (Armstrong et al. 2012: 305). For starters, the public would require sufficient comprehension of the pertinent ethical issues—issues that are complicated and lack either direct or analogous empirical examples to build a framework of moral judgement. Goertzel and Bugaj state: "an intelligent agent should not be expected to act according to an ethical principle unless there are many examples of the principle-in-action in its own direct or observational experience" and should act "in a way that others can understand the principles underlying ones actions" while also "having others directly imitate ones actions in directly comparable contexts" (2008: 379). I consider this framework problematic, since an AGI's approach must be partly inscrutable, because it is superintelligent with respect to all humans. The notion of public engagement in AGI governance without useful examples and sufficient understanding to guide them would therefore increase noise over signal.

From the Luddites to Clinton-era anti-trust campaigns, technician disciplines have a litany of historical examples where the public resisted rapid technological change through fear of having their livelihoods replaced by machines and technocrats: "humans express a pathology of irrational behavior after the uncomfortable threatening of their mental supremacy. The fear is not unjustified if one considers current technological advancements" (Galanos 2017: 574). The recent hostility towards technology firms with a significant stake in AI development (e.g., Google, Facebook) illustrates a continuation of this trend. Yet, despite the hostility, there is a disturbing absence today of robust national and international AI regulations. American government today acts on AI in response to public fear, and similarly "shortsighted decisions of upper management" (Turchin and Denkenberger 2018: 153) for AGI may prove disastrous. Technicians therefore do not blindly endorse top–down governance, but instead argue for giving top–down authority to *technical experts*. Rather than being anti-democratic per se, technicians fear that in the particular case of AGI, over-emphasizing layperson concerns might continue a pattern whereby their fears stimulate electorally minded politicians to govern myopically, given that they too are largely technologically illiterate and respond mainly to short-term incentives. As Montes and Goertzel state: "we understand some of the limitations to democracy… [we advocate] carefully balance decentralized democratic governance with a limited degree of high-level benevolent stewardship" (2019: 357).

I see the democracy vs. technocracy discussion as a debate about what *kinds of experts* are necessary for AGI governance. As outlined above, technicians feel cloistered decision-making by election-minded representatives and their staff bureaucrats is the real hindrance to good top–down technology policy-making, not technocrats: "technologists and scientists have a "distinctive-competence" in these kinds of ethical activities… scientists and technologists, due to their high intelligence, are well equipped to publicize well-informed and persuasive justifications" (Singer 2015: 3). They advocate a heightened role for the *right kind of expert*, as described by Singer: "The prospect of the rise of the scientific-middle as a force for good… holds out the promise of some moral progress. There is currently a great need to inject scientific and ethical habits of thought into the entire global distributed governance process" (Singer 2015: 10).

Technicians suggest how to bring appropriate experts into the fold. One lesson that was highlighted was nuclear weapons development during the Manhattan Project. Instead of the government-controlled funding and explicitly militaristic ambitions characterizing that project, for AGI, they felt that "modest amounts of funding should be distributed to a variety of groups following different approaches, instead of large amounts of funding being given to a "Manhattan Project" type crash program following one approach." (Baum et al. 2011: 190). This method mitigates the risk of an uninformed-yet-powerful regulator, removes a winner-takes-all dynamic to fund-seeking, and incentivizes collaboration. Technicians also sought increasingly multidisciplinary intellectual and professional contributors. This included research from currently marginal academic disciplines like sociology, psychology, art, and biology: "sociology and other social science need to acquire an adequate understanding of how artificial intelligence is and should be grown into a social actor" (Mlynar et al. 2018: 131), specialized philosophical and ethical education for applied disciplines "in the context of… artificial intelligence safety research" (Yampolskiy 2015: 236), and effective collaborations "among academia, industry, engineers and policy-makers about the foundations of morality and ethics in regards of A(G)I" (Dameski 2018: 43).

I find that PADs are more democratic in their approach. One justification comes from a different reading of the history of technology and inequality: "the history of technological diffusion suggests that there is no reason to assume that all will benefit equally" (Brundage 2014: 368). While technicians effectively stressed a kind of constructive inequality, PADs were reluctant to consider this a suitable justification for outsized power and influence. They did not see power as something technicians are intrinsically better at handling responsibly. Bostrom outlines a scenario where a well-controlled AGI could be built so as to not be "generically philanthropic but would instead give it the more limited goal of serving only some small group" (2003c: 4). A survey conducted by PADs further illustrates this concern, where "a notable difference between the 'theoretical' and the 'technical' groups" (Müller and Bostrom 2016: 12) was observed. More from the 'theoretical' group (here approximating PADs) felt AGI was likely to create extremely bad outcomes (21% of respondents) than the 'technical' group (here approximating technicians, 7%). If PADs are less convinced that AGI will be 'good', then the regulatory virtues of the technicians who build it are less relevant.

PADs pushed back against the 'right kind of expert' argument by questioning whether AGI expertise meaningfully exists: "[AGI] could mark the end of science as such, since acquiring a complete explanatory-predictive 'theory of everything' would likely constitute an instrumental value for any given superintelligence with any set of final goals… the first superintelligence could mark the end of history in a very significant sense" (Torres 2018: 357). In scenarios like those outlined by Torres, can any human being claim to be 'an AGI expert'? Is the implicit claim of expertise—authority and knowledge society can trust—really appropriate to AGI? Expertise might prove be illusory, or at least trivial, once AGI is operative. In fact, some PADs believe that this 'illusory expertise' is already a problem in present-day AI. For instance: "with AI predictions being considered more uncertain than even the 'softest' sciences… efforts should be made to connect general prediction with some near-term empirical evidence" (Armstrong et al. 2014: 323). There is essentially *no empirical evidence*, near-term or otherwise, to make AGI predictions and plan accordingly. They argue that the difference between expert and lay judgement is slight, for when considering: "a machine that has not yet been built, and for which a detailed plan does not exist, there is little opportunity for the hypothesis-prediction-testing cycle" (2014: 323). This explains why PADs endorsed slower, crowdsourced, participatory modes of deliberation. Returning to the survey by Müller and Bostrom (2016), the wide range of expert AGI predictions supports the philosopher's case for their limited utility and questionable accuracy.

The contrasts I find between technicians and PADs align with my expectations. Philosopher circumspection toward expert-led AGI regulation accords with an overwhelming endorsement in the wider philosophical community for egalitarian political arrangements (Bourget and Chalmers 2014). Since they place greater value on public input, I note that a recent National Science Foundation (2018) survey found the American public held a generally favorable view of science and technology, but they were concerned by the accelerated rate of change and the increase of 'black-boxes' that distance technologists from the public who use their creations. Yet, technicians can point to expert commentators who highlight declining levels of public financing, increased regulatory overreach, and developmental stagnation in science and technology as being *the* critical problem of modern industrialized society (see Cowen 2011; Thiel 2014; Gruber and Johnson 2019). Bostrom, widely regarded as the pre-eminent voice on AGI, has deemed AGI "philosophy with a deadline" (2014: 314), and "suggests that philosophic progress can be maximized via an indirect path rather than by immediate philosophizing… we could postpone work on some of the eternal questions for a little while, delegating that task to our hopefully more competent successors—in order to focus on a more pressing challenge: increasing the chance that we will actually have competent successors" (315). Both PADs and technicians are united by a meta-admission: current expertise is far from sufficient.

5.2 Deontology vs. utilitarianism

There are a number of normative ethical theories in contemporary academia. Two systems, thought to be contrasting and arguably the most widely used, are utilitarianism and deontology. I find technicians use deontological thinking more than PADs, while both thoroughly critique utilitarianism. Briefly, normative ethics considers what standards and principals of moral thought we should use to do more right and less wrong. Deontology prioritizes adherence to a set of rules and/or duties when acting in the world. According to Kant, deontology's intellectual bellwether, this amounts to the injunction that agents "act only according to that maxim whereby you can, at the same time, will that it should become a universal law" (1993). Utilitarianism, sometimes referred to as consequentialism (strictly speaking, utilitarianism is a specific form of consequentialism), is the brainchild of Jeremy Bentham, who saw "two sovereign masters" of mankind, "pain and pleasure" (Bentham 1789). Utilitarianism judges moral action via its consequences, with the morally right being that that brings about the most pleasure, good or happiness, and vice versa. Other systems like virtue ethics and stoicism received fleeting attention in my corpus, and are not discussed here.

Technicians Goertzel and Bugaj endorsed a deontological approach as it affords reversibility: "where a moral act within a particular situation is evaluated in terms of whether or not the act would be satisfactory even if particular persons were to switch roles within the situation" (2008: 368). This approach means moral rules can be assessed from multiple perspectives, by both humans and machines, encouraging a process of convergence and refinement on a set of ethical principles and their good-faith interpretation. These scholars suggest that a super-rational AGI system will inevitably come to realize this approach: "we suggest that once the capability for formal reasoning matures, the categorical imperative and the quest for logical coherence naturally emerge" (Goertzel and Bugaj 2008: 376). A deontological rule set of this kind could form a 'baseline' for normative AGI ethics that avoids black swan scenarios: "we are not interested in a consistent and complete system of ethics that will tell us in advance what we ought to do in any given circumstance, we are only interested in guidelines that are strong enough to stave off existential threat" (Kornai 2014: 423). Some stressed "the urgent need to adopt the 23 Asilomar Principles" (Wogu et al. 2017), currently the only politically substantive effort at AGI policy-making (the state of California has adopted them), which are thoroughly deontological in nature.

Yet, technicians noted the limits to a deontological approach also, namely instances where rules contradict or undermine other rules: "ethical rule sets do not work well in situations where all possible actions have negative consequences… [and] as the ethical rule set increases to cover the widening set of circumstances, it become sever more challenging to avoid internal conflict and ensure consistency between the rules" (Rolf and Crook 2016: 3). Some suggested rules are necessary but not sufficient, and can only be effective if we avoid relying on them as sole means of encoding ethical axioms: "no set of rules can ever capture every possible situation, and that the interaction of rules may lead to unforeseen circumstances and undetectable loopholes" (Yampolskiy 2015: 125–6). Doubts were also raised about the robustness of a rules-based approach if AGI becomes conscious, given that conscious agents (like humans) are able to make and alter rules if sufficiently intelligent: "each time an application of a rule comes to consciousness, it, like every conscious event, becomes subject to perceptual learning" (Wallach et al. 2010: 475).

Technicians criticize utilitarianism, because it "considers very good and very bad scenarios to be symmetrical" (Sotala and Gloor 2017: 391). This is precisely the opposite of the 'black swan' logic discussed in my existential risk section, where it was agreed that the negative consequences of AGI far outweigh the positive ones. A utilitarian calculus requires reliable information on which to judge expected consequences and their moral value—technicians doubted that such information could be relied upon, with inevitable "choices where information is incomplete or of questionable accuracy, or where the consequences of possible courses of action cannot be known in advance" (Wallach et al. 2011). They also express metaphysical concerns, namely whether an AGI will have the same (or any) metaphysical assumptions as humans when they assess moral consequences: "a superintelligent being may have very good reasons to deny some of the commonsensical assumptions about space and time, actions and consequences, goals and purposes…" (Kornai 2014: 426).

The philosopher's critique of deontology was more extensive, and thematically different, to the technicians. They fear the possibility of 'perverse instantiation' of rules in AGI, whereby the given rule is adhered to technically but not in intended meaning or spirit: "attempts to spell out such constraints [rules], or suitably prioritise them, may lead to perverse outcomes in certain situations… the absence of one particular consideration in a moral system could lead to dangerous results even if the rest of the system is well-developed" (Brundage 2014: 363). A much simpler critique notes that intelligence allows an agent to amend or ignore rules, especially when they are explicit and formal: "AIs which are smarter than humans… can bypass these rules, if they so choose" (Yampolskiy and Fox 2013: 4). In fact, PADs noted how the very nature of rules is to simultaneously create loopholes, and "any slightly intelligent machine will discover all the loopholes in our legal, economic and ethical systems as well or better as human beings" (Yampolskiy 2013: 407), leading Yampolskiy to conclude that "explicit rules are easy to implement, but are unlikely to serve the intended purpose" (2013: 408). Rules, according to Yampolskiy, are problematic without a rich understanding of context. In another article, he and co-author Joshua Fox state that "whatever the rules imposed, it would be dangerous to constrain the behavior of advanced artificial intelligences which interpret these rules without regard for the complex ensemble of human values" (2013: 4).

The philosopher's critique of utilitarianism was similar to technicians. It centered around what, and how, to value future situations or states to consider them morally good or bad. For instance, if we instruct AGI to maximize a certain good, the pursuit of other less-good states could suffer when maximally pursuing this 'better' one. This "illuminates a general concern within utilitarian approaches to machine ethics—that the maximizing, monistic nature of such ethical frameworks may justify dangerous actions on a large scale" (Brundage 2014: 362). PADs too were concerned by the implied symmetry between 'good' and 'bad', saying that this assumption runs counter-valent to a human tendency to prioritize moral risk-aversion: "assume you may press a button such that with probability 0.5 a random person's leg will be broken, and with probability 0.5 someone's broken leg will be healed. I think it goes without saying that it is immoral to press the button" (Beckers 2017: 92). Finally, I find that the open question of whether AGI could be conscious made PADs cautious of the utilitarian approach. A conscious AGI might be able to generate subsequent subjective mental states, minds, or consciousnesses. Since these may be capable of joy, suffering, and other morally significant experiences, employing a utilitarian calculus accurately could prove impossible: "emulations can rapidly produce a large amount of positive of negative value if they are in extreme states: they might count for more in utilitarian calculations. Does human emulation have a right to real-time?" (Sandberg 2014: 451).

To summarize, discourse on normative ethics differed from my expectations. Some long-standing critiques of utilitarian thought take the pain vs. pleasure calculus to its logical extreme to illustrate fundamental and inevitable flaws if it is universally adopted (see Arrow 1951 or Nozick 1974 for classic examples). Economist Arrow (1951), for instance, suggests that comparing utilities must become incoherent given enough choices, even in an agent capable of appropriating all of the choices individually (1951; see also Nozick 1974 for a similar approach). AGI, as currently imagined, fits many of the same criteria: an agent with functionally infinite and uncertain dimensions to its experience, decision-making capacity, intentions, and values. It could be an agent capable of pure utility maximization, yet according to Arrow, ethical decision-making would still become incoherent. When added to other concerns in the data about the disputed role of time, the difficulty of prioritizing the good against the bad, and the uncertain relationship between stating a desirable outcome and actually pursuing or realizing it, I find that utilitarianism is deemed an inappropriate normative framework for AGI.

Regarding deontology, I note that algorithmic computation is intrinsically rules-based. Therefore, there may exist a more natural affinity between deontology and the *modus operandi* of technicians than PADs. It is typical of technician disciplines to consider physical systems as fundamentally adhering to some basic set of rules. Machines are valued by society precisely, because they rigorously and repeatably adhere to a basic rule structure—those that do this poorly are considered substandard or broken. As technicians endorsed a basic rule structure for AGI that guards against worst-case moral scenarios, I suggest that this is because in critical applications, this is what any computational or machine-entity can be best-relied upon to execute. While utilitarianism has rules as well (namely, 'maximize utility'), the metric of success refers to outcomes. Deontology privileges rules over outcomes to some greater degree than utilitarianism—if consistently bad outcomes happen as a result of some deontological rule, it could eventually be that the rule is considered wrong. However, in utilitarianism, *any* net-negative outcomes violate the only rule—maximize utility. The rule itself is never in question, whereas deontology is essentially judged by the rules it embodies. This mirrors machines in that if they adhere to the rules they were designed for yet produce consistently bad outcomes, it is the fault of the rule-makers, i.e., the designers or, for AGI, the technicians. Also, the 'reversible' approach of Goertzel and Bugaj (2008) was reminiscent of John Rawls 'Veil of Ignorance', which has proven a very robust, popular deontological foundation for modern society. PADs were essentially concerned that AGI may bypass or alter any rules it is instructed to obey. Their focus is over a longer time-horizon than technicians; they were disinterested in pragmatic concerns of what works best in the present day. The wider epistemological lens of philosophy, which seeks "to understand how things in the broadest possible sense of the term hang together in the broadest possible sense of the term" (Sellars 1962: 35), offers one explanation for this finding.

5.3 The possibility of moral realism

I find technicians tend to believe that an AGI could uncover and verify 'moral facts' more than PADs. Whether there are 'moral facts' is a long-standing philosophical debate—does morality have inviolable axioms or is it ultimately relativistic? The significance of this question can be summarized as follows—if moral reality exists, then presumably a discreet set of ethical principles can be derived by, and codified into, an AGI. One-way technicians justified the moral realist position was noting the presence of converging ethical values in the natural world, arrived at via a compounding of moral decisions by a large number of agents: "most moral systems created by systems that enter into similar moral scenarios (i.e., human collectives) are alike, and universality or widespread adoption in some basic moral rules can be discovered throughout them. Examples in human moral systems are of the immorality of murder, rape, sexual acts with children, incest, lying, irresponsible or unnecessary disturbance or damage, or similar" (Dameski 2018: 49). A sociological variant of this claim from Montes and Goertzel notes that "basic value structures have shown common trends across individuals and societies" (2019: 357). Some suggested we view this process as part of biological evolution. Adopting this lens, ethical precepts are arrived at through forces of selection and fitness, becoming incorporated into the generic human cognitive and biological architecture as they promote survival in varied environments over time. This culminates in a set of moral universals: "in a very real sense, "good" and "bad" actually generate seemingly physical sensations that have evolved to help us survive" (Waser 2011: 160). Finally, some technicians believed a similar convergence happens in the ethics of emerging technologies, specifically: "almost every [moral] claim made about one specific technology… can be literally duplicated with reference to any of the other technologies… without any significant loss of moral or scientific support" (Singer 2015: 2). These trends legitimated tasking AGI with moral fact-finding.

PADs were more skeptical, concerned about dangerous or unintended consequences from an AGI seeking out moral facts. Realistically, humans cannot know whether this is a coherent goal in advance of instructing the AGI to aim for it. For example, Worley states: "if we suppose [moral] realism, then we could build aligned AGI on the presupposition that it could at least discover moral facts even if no moral facts were specified in advance… now suppose this assumption is false and that moral facts do not exist… then our moral-facts-assuming AGI would either never discover any moral facts to guide its behavior when human values are in conflict or would assume arbitrary moral propositions to be facts…" (2018, 4–5). In Worley's scenario, an AGI proclaiming to have found a 'moral fact' could result in the totalitarian enforcement of inappropriate ethical mandates. Other PADs believe that moral fact-finding is the wrong goal for AGI. Instead, moral realism should be eschewed for a more pragmatic goal—*improving* moral and ethical thinking. If moral facts do exist, then AGI may arrive at them via this more modest goal. For example, "*to the extent* that ethics is a cognitive pursuit, a superintelligence could do it better than humans… questions about ethics, *in so far as they have correct answers* that can be arrived at by reasoning and weighting of evidence, could be more accurately answered by superintelligence than by humans" (Bostrom 2003c: 3, emphasis added). Since we cannot know whether optimal ethical action can be arrived at purely cognitive processes, not setting AGI the goal of moral fact-finding is prudent to PADs.

Epistemology can explain the difference between the two groups. Technician claims are mechanistic and draw their energy from empirical observations. They extrapolate empirical trends—evolution, human societies, technology—to tell a causal story about how moral facts can be ascertained within any group of agents. But are these trends specific to human and/or biological systems? Are these trends set to continue? I believe you must adopt a computational view of cognition to extend these trends to machines. Many technical fields adopt the 'mind as computer' perspective (see Rescorla 2020) or, further still, consider the universe fundamentally a physical system enacting computations (e.g., Lloyd 2000). If you operate under these assumptions, transposing moral fact-finding processes from man to machine is plausible. However, this says nothing about whether trend of moral convergence will continue. Here, I observe how similar the procedure of establishing patterns from past empirical data, imposing theoretical explanations, and then casually extrapolating them to make future predictions, is reminiscent of the logic of machine learning, the current state-of-the-art in AI. This affirms my prior expectations: technicians are inclined to consider moral reality a mechanistic problem that can be solved by applying principles of the natural world, themselves derived through analysis of empirical data. By contrast, epistemological uncertainty is where philosophy thrives. PADs are characterized by their capacity to scrutinize and be skeptical of seemingly definitive conclusions. Despite moral realism being the majority meta-ethical stance of professional philosophers (Bourget and Chalmers 2014), there is no philosophical means to establishing relative certainty, so their epistemological commitments demand caution. Nevertheless, in a binary choice, I speculate that PADs would join technicians to predict that AGI will one day unearth moral facts. Their shared optimism for ethical *progress* using AGI indicates that computational, mechanistic, empirical approaches to doing ethics are at least effective.

6 Discussion and conclusion

The AGI ethics debate is primarily concerned with mitigating existential risk—technicians and PADs both agree on this. Both groups confidently predict AGI will be the most ethically consequential technology ever created. Accordingly, a suitably proactive response demands greater funding, research and input from the academy, governments, and the private sector. Technicians and PADs both endorse a highly precautionary approach, deemphasizing the moral and humanitarian well-being an AGI could provide to first focus on preventing worst-case scenarios. Addressing these black swans has stimulated a demand for better knowledge in two adjacent research areas. First, a model of what consciousness is, a theory of its development, and a sense of whether non-organic entities or systems can become conscious. Second, teasing out broader similarities and differences between organic and mechanical systems, so the utility, generalizability, and predictive power of empirical trends can be better assessed. The case of AGI highlights agreement between PADs and technicians on humanity's extensive moral obligation towards future generations. In an age of multiple existential risks, including those unrelated to AGI, e.g., climate change, humans today bear direct responsibility for minimizing either termination or increased suffering of future lives.

A finding that defied my expectations was how technicians embraced deontological normative ethics more fully than PADs. I suggest that technicians' epistemological commitments to reductionism, and the foundational importance of rules in computation, as key reasons for this. In some sense, a computer is successful to the extent that it follows a set of predefined rules across as wide an array of contexts as possible. By contrast, utilitarianism underwent extensive criticism by both groups. I propose this is because AGI functions as a case equivalent to the extreme fictional agents used in robust and long-standing anti-utilitarian thought experiments. On the topic of AGI governance, technicians believe a surplus of democratic input could complicate or burden responses to AGI-related issues, mainly by providing bureaucrats and politicians with top–down control better utilized by technicians themselves. It also follows that technicians are more liable to consider technical or procedural creations as 'good', since they understand them better, and have greater personal and professional investment in their success. This would incline them towards technocracy, as it implies focusing upon improving an AGI whose existence in the world is already taken as positive or necessary. Notable efforts at internal ethical regulation and reporting standards by organizations like the American Association of Artificial Intelligence and the Conference on Neural Information Processing Systems have been developed. While these safeguards are encouraging, they are nevertheless administered and judged by the same community. PADs rejected this perspective, doubting whether technicians would in fact do a significantly better job. Finally, while both groups believe AGI will speed progress on the overall project of ethics, technicians felt that it would likely be able to explicitly formulate moral facts, a position that PADs were highly skeptical of. Technicians, being steeped in the language of computation and quantity, are liable to view morality similarly—as a formal problem with a derivable solution.

It is notable how quickly this debate is advancing, as AGI comes closer to being realized. As AI practitioners wrestle with their increasing impact on the world, they are increasingly turning their attention towards ethics. Public and private AI organizations are establishing ethics boards, and agreements like the Asilomar Principles are receiving increased attention as possible governance strategies. The public is increasingly aware of the ethical problems posed by AI. Prominent figures are talking increasingly openly about the problems of AGI, also. This has caused AI practitioners to place increased premium on transparency, and to think through the consequences of their technologies more closely. This public debate will inevitably advance to AGI ethics as the urgency of this technology becomes more apparent. AGI ethics is a crucial humanitarian issue.

References

Aliman N-M (2017) Malevolent cyborgization. Paper presented at the 10th international conference on artificial general intelligence, AGI 2017, August 15, 2017–August 18, 2017, Melbourne, VIC, Australia.

Armstrong S, Sandberg A, Bostrom N (2012) Thinking inside the box: controlling and using an Oracle AI. Mind Mach 22(4):299–324

Armstrong S, Sotala K, hÉigeartaigh SSO (2014) The errors, insights and lessons of famous AI predictions—and what they mean for the future. J Exp Theor Artif in 26(3):317–342

Arrow KJ (1951) Social choice and individual values. Wiley

Baum SD, Goertzel B, Goertzel TG (2011) How long until human-level AI? Results from an expert assessment. Technol Forcast Soc Change 78(1):185–195

Beckers S (2017) AAAI: an argument against artificial intelligence. In: Müller V (ed) Philosophy and theory of artificial intelligence 2017. Springer, pp 235–247

Beckstead N (2013) On the overwhelming importance of shaping the far future. Doctoral dissertation, Rutgers University-Graduate School-New Brunswick

Bentham J (1789) An introduction to the principles of morals. Athlone

Björnsson G (2012) Do 'objectivist' features of moral discourse and thinking support moral objectivism? J Ethics 16(4):367–393

Borrego Á, Anglada L (2016) Faculty information behaviour in the electronic environment. New Libr World 117:173–185

Bostrom N (2003a) Astronomical waste: the opportunity cost of delayed technological development. Utilitas 15(3):308–314

Bostrom N (2003b) When machines outsmart humans. Futures 35(7):759–764

Bostrom N (2003c) Ethical issues in advanced artificial intelligence. Science fiction and philosophy: from time travel to superintelligence. Wiley, pp 277–284

Bostrom N (2003d) Are we living in a computer simulation? Philos Q 53(211):243–255

Bostrom N (2014) Superintelligence: paths, dangers, strategies. Oxford University Press

Bourget D, Chalmers DJ (2014) What do philosophers believe? Philos Stud 170(3):465–500

Brooks RA (2017) Robotics pioneer Rodney Brooks debunks AI hype seven ways. MIT technology review. https://www.technologyreview.com/s/609048/the-seven-deadly-sins-of-ai-predictions/. Accessed 3 Oct 2021

Brundage M (2014) Limitations and risks of machine ethics. J Exp Theor Artif Intell 26(3):355–372

Collins H (2018) Artifictional intelligence: against humanity's surrender to computers. Wiley

Cortese FAB (2014) The maximally distributed intelligence explosion. In: Paper presented at the 2014 AAAI spring symposium series

Cowen T (2011) The great stagnation. Dutton & Co

Cowen T (2018) Stubborn attachments. Stripe Press

Dameski A (2018) A comprehensive ethical framework for AI entities: foundations. In: Paper presented at the artificial general intelligence. 11th international conference, AGI 2018, 22–25 Aug 2018, Cham, Switzerland

Danaher J (2015) Why AI doomsayers are like sceptical theists and why it matters. Mind Mach 25(3):231–246

Engineering Village (2017) Engineering Village fact sheet. https://www.elsevier.com/__data/assets/pdf_file/0008/314693/EV_Factsheet_-Engineering-Village-Databases_July-2017.pdf. Accessed 16 Feb 2019

Elo S, Kyngäs H (2008) The qualitative content analysis process. J Adv Nurs 62(1):107–115

Etzioni O (2016) No, the experts don't think superintelligent AI is a threat to humanity. MIT technology review. https://www.technologyreview.com/2016/09/20/70131/no-the-experts-dont-think-superintelligent-ai-is-a-threat-to-humanity/. Accessed 23 Aug 2019

Evans JH (2002) Playing god?: human genetic engineering and the rationalization of public bioethical debate. University of Chicago Press

Evans JH (2006) Between technocracy and democratic legitimation: a proposed compromise position for common morality public bioethics. J Med Philos 31(3):213–234

Everitt T, Lea G, Hutter M (2018) AGI safety literature review. Preprint http://arXiv.org/abs/1805.01109.

Fieser J (2021) Ethics. The internet encyclopedia of philosophy, ISSN 2161-0002. https://www.iep.utm.edu/. Accessed 11 Mar 2021

Fitzgerald M, Boddy A, Baum SD (2020) Survey of artificial general intelligence projects for ethics, risk, and policy: technical report 20-1. Global Catastrophic Risk Insitute

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

Freed S (2020) AGI needs the humanities. In: International conference on artificial general intelligence. Springer, Cham, pp 107–115

Galanos V (2017) Singularitarianism and schizophrenia. AI Soc 32(4):573–590

Goertzel B, Bugaj SV (2008) Stages of ethical development in artificial general intelligence systems. In: Paper presented at the artificial general intelligence 2008. Proceedings of the first AGI conference.

Good IJ (1965) Speculations concerning the first ultraintelligent machine advances in computers. Academic Press

Grace K, Salvatier J, Dafoe A, Zhang B, Evans O (2017) When will AI exceed human performance? Evidence from AI experts. Preprint http://arXiv.org/abs/1705.08807.

Graham R, Compton J, Meador K (2019) A systematic review of peer-reviewed literature authored by medical professionals regarding US biomedicine's role in responding to climate change. Prev Med Rep 13:132–138

Greaves H (2017) Population axiology. Philos Compass 12(11):e12442

Gruber J, Johnson S (2019) Jump-starting America. Public Affairs

Hawken P (2017) Drawdown. Penguin Random House

Irving G, Askell A (2019) AI safety needs social scientists. Distill 4(2):e14

Kahneman D (2011) Thinking, fast and slow. Macmillan

Kant I (1993) Grounding for the metaphysics of morals. Ellington (**Translated by J. W. Hackett, 1975**)

Koch C (2019) The feeling of life itself. MIT Press

Kornai A (2014) Bounding the impact of AGI. J Exp Theor Artif Intell 26(3):417–438

Kurzweil R (2005) The singularity is near. Gerald Duckworth & Co

Lloyd S (2000) Ultimate physical limits to computation. Nature 406(6799):1047–1054

Mlynar J, Alavi HS, Verma H, Cantoni L (2018) Towards a sociological conception of artificial intelligence. In: Paper presented at the artificial general intelligence. 11th international conference, AGI 2018, 22–25 Aug 2018, Cham, Switzerland.

Montes GA, Goertzel B (2019) Distributed, decentralized, and democratized artificial intelligence. Technol Forecast Soc 141:354–358

Müller VC, Bostrom N (2016) Future progress in artificial intelligence: a survey of expert opinion. Fundamental issues of artificial intelligence. Springer, pp 555–572

National Science Board (2018) Science and technology indicators, 2018. https://www.nsf.gov/statistics/2018/nsb20181/assets/404/science-and-technology-public-attitudes-and-understanding.pdf. Accessed 9 Sept 2019

Nozick R (1974) Anarchy, state, and utopia. Basic Books

Ord T (2020) The precipice: existential risk and the future of humanity. Hachette Books

PhilPapers (2017) About PhilPapers. https://philpapers.org/help/about.html. Accessed 12 Feb 2019

Pölzler T, Wright JC (2019) Empirical research on folk moral objectivism. Philos Compass 14(5):e12589

Prunkl C, Whittlestone J (2020) Beyond near- and long-term. In: Proceedings of the AAAI/ACM conference on AI, ethics, and society. http://dx.doi.org/10.1145/3375627.3375803

Rescorla M (2020) The computational theory of mind. In: Zalta EN (eds) The Stanford encyclopedia of philosophy (Spring 2020 edition). https://plato.stanford.edu/archives/spr2020/entries/computational-mind/. Accessed 13 Apr 2020

Rolf M, Crook N (2016) What if: robots create novel goals? Ethics based on social value systems. In: EDIA @ ECAI, pp 20–25

Russel S (2019) Human compatible. Viking Press

Sandberg A (2014) Ethics of brain emulations. J Exp Theor Artif Intell 26(3):439–457

Sandberg A, Bostrom N (2008) Global catastrophic risks survey. Technical report #2008-1, Future of Humanity Institute, Oxford University, pp 1-5

Sayre-McCord G (2015) Moral realism. In: Zalta EN (eds) The Stanford encyclopedia of philosophy (Winter 2020 edition). https://plato.stanford.edu/archives/win2020/entries/moral-realism/. Accessed 12 Mar 2021

Sellars W (1962) Philosophy and the scientific image of man. In: Colodny R (ed) Frontiers of science and philosophy. University of Pittsburgh Press, pp 369–408

Singer AE (2015) Stakeholder capitalism and convergent technologies. Int J Soc Org Dyn 4(2):1–11

Snow CP (1959) The two cultures. New Statesman 6:413–414

Sotala K, Gloor L (2017) Superintelligence as a cause or cure for risks of astronomical suffering. Informatica 41(4):389–400

Strauss A, Corbin JM (1997) Grounded theory in practice. Sage

Taleb NN (2007) The black swan. Random House

Tegmark M (2017) Life 3.0. Knopf

Thiel P (2014) Zero to one. Crown Business Press

Torres P (2018) Superintelligence and the future of governance: on prioritizing the control problem at the end of history. In: Yampolskiy R (ed) Artificial intelligence safety and security. CRC Press, pp 357–374

Totschnig W (2017) The problem of superintelligence: political, not technological. AI Soc 34(4):907–920

Turchin A, Denkenberger D (2018) Classification of global catastrophic risks connected with artificial intelligence. AI Soc 35(1):147–163

Turing A (1950) Computing machinery and intelligence. Mind 49:433–460

Wallach W, Franklin S, Allen C (2010) A conceptual and computational model of moral decision making in human and artificial agents. Top Cogn Sci 2(3):454–485

Wallach W, Allen C, Franklin S (2011) Consciousness and ethics: artificially conscious moral agents. Int J Mach Conscious 3(1):177–192

Walsh T (2018) Expert and non-expert opinion about technological unemployment. Int J Autom Comput 15(5):637–642

Waser M (2011) Rational universal benevolence: simpler, safer, and wiser than “friendly AI”. In: Paper presented at the artificial general intelligence. 4th international conference, AGI 2011, proceedings

Wogu IAP, Olu-Owolabi FE, Assibong PA, Agoha et al. (2017) Artificial intelligence, alienation and ontological problems of other minds: a critical investigation into the future of man and machines. In: 2017 international conference on computing networking and informatics (ICCNI). IEEE, pp 1–10

Worley GG (2018) Robustness to fundamental uncertainty in AGI alignment. Preprint http://arXiv.org/abs/1807.09836

Yampolskiy RV (2013) What to do with the singularity paradox? In: Muller V (ed) Philosophy and theory of artificial intelligence. Springer, pp 397–413

Yampolskiy RV (2015) Artificial superintelligence: a futuristic approach. Chapman and Hall

Yampolskiy R, Fox J (2013) Safety engineering for artificial general intelligence. Topoi 32(2):217–226

Yetisen AK (2018) Biohacking. Trends Biotechnol 36(8):744–747

Yudkowsky E (2008) Artificial intelligence as a positive and negative factor in global risk. In: Bostrom N, Cirkovic M (eds) Global catastrophic risks. Oxford University Press, Oxford

Zhang B, Dafoe A (2019) Artificial intelligence: American attitudes and trends. Available at https://ssrn.com/abstract=3312874. Accessed 21 June 2019