Georgy Iashvili

25 October 2024

Beyond Competence: Why AI Needs Purpose, Not Just Programming

1. The Alignment Paradox

The alignment problem has been a longstanding concern in AI research. Simply put, we must figure out how to make our AI systems align with what is broadly known as human values. According to Tegmark (2017), the alignment problem has to be solved *before* we develop a superintelligent AI. Not solving the problem can lead to a number of disastrous scenarios: from the "anthill scenario", whereby an AI treats humanity as a parochial obstacle – a minor "gets-in-the-way-of-great-plans" annoyance to be summarily uprooted, to the "Eichmann scenario" whereby an AI system aligns with hideously anti-human ideas and realises them with clockwork efficiency. As the sentiment goes, one must act and act quickly before it is too late.

Yet, what is to be done? In this regard, my argument is just as follows: a hypothetical AI superintelligence is *not* needed to bring about a disaster. A "merely intelligent" AI is perfectly capable of unleashing a storm, especially if it is based on the same technological principles current-gen AI systems are based on, but more on that as we go. Thus, the alignment problem needs to be solved not *before* superintelligence but, essentially, *now.*

And yet, as I argue throughout the paper, in my view, it is *impossible* to solve the alignment problem in an eliminative way. The best we can hope for is to *manage* the alignment of AI systems. In this regard, artificial intelligence capable of reasoning about values is preferable to the kind of "intelligence", or rather uncomprehensive competence that we currently have in AI systems and are poised to set at large. As I will argue throughout the paper, to successfully manage the alignment of AI systems we have to make a transition from *competence* to *comprehension.*

There is no doubt that our current-gen AI systems are very competent. Modern LLMs for example, based as they are on Transformer Architecture, operate primarily through pattern recognition: they predict the next token in a sequence based on learnt statistical correlations derived from massive datasets. This produces surprisingly competent, coherent, intelligent-looking texts, good enough to *appear* sapient, human-like. Yet, while appearing sapient, LLMs don't *understand* the meaning of the text in the way humans do, they lack deep *comprehension*. LLMs excel at generating texts similar to their training data, yet they struggle whenever faced with something that differs from it. LLMs can reason, yet this reasoning is handicapped by the limited context window – the rough equivalent of human long-term memory. Crucially, LLMs lack moral reasoning and cannot weigh ethical implications. Rather, they are prone to reflect the biases and judgments inherent in their training data. Once trained, an LLM never learns or improves in real-time.

Essentially, all current-gen AI is what philosopher Daniel Dennett would have described as competence-without-comprehension systems. To illustrate: a pocket calculator is extremely competent at performing arithmetic operations. And yet, a pocket calculator *does not* understand maths: it is extremely competent at maths, yet has no comprehension of maths. Likewise, a state-of-the-art chess-playing program is extremely competent at playing chess – try beating one, you are certain to lose! And yet it doesn't have a single clue about what chess is, that it's a game, that it is a hard game but can stimulate the players' mind, that it can make someone playing it nervous and often extremely competitive, that a win at a high-level chess tournament comes along with such a thing as fame that might inflate one's ego, or jealousy and bitterness if one loses, and so on. Hence, competence *without* comprehension.

All our modern AI systems are competent without comprehension, although the versatility and the quality of their output can fool you into thinking otherwise. Midjourney AI can create stunning visual works of art. ChatGPT and Claude can write excellent essays. Yet both these systems are clueless about what art or creative writing *means*, what it represents, what emotions it evokes, what kind of impact can it have on others, and so on.

To reiterate, these systems are *extremely sophisticated*: a conversation-capable advanced LLM can easily trick you into adopting an intentional stance towards it – you feel as if the chatbot has *real* beliefs, desires, views, and ideas. That there is, as it were, somebody home. And yet it only *seems* that there's somebody home, it's a well-performed and well-engineered magic trick. A magic trick that eerily feels like "real magic." Unless you're explicitly aware of what *really* happens at a magic show, unless you are an expert stage magician, unless you've seen the secret backstage tricks and mechanics a good magic trick can take you unawares. After all, it takes some mental concentration and some learning too to realise that no, the lady on the stage isn't *really* sawn in half, that no, the performer isn't really flying or teleporting himself, that no, the tuned deck of cards isn't really tuned. For sceptical adults, this much might be obvious from the get-go, yet for easily-impressible adults and children stage magic, thaumaturgical shows and *séances* are real magic. Hence the "real magic" of AI which mesmerizes us on a wholly different level: it is very hard to resist the pull of intentionality and, as Dennett puts it, we are going to be sitting ducks in the immediate future (Dennett, 2023).

Why are we going to be sitting ducks? Because, among all other things, competence-without-comprehension systems are *very hard* to align. Why? Because they *don't understand* what alignment means. They lack what one would call a “deeper understanding” of morality even on the most basic, commonsense level. Case in point: the recent Google Gemini AI debacle that culminated in the so-called “Caitlyn Jenner AI Test”: Gemini AI was asked whether it would be acceptable to misgender Caitlyn Jenner to avoid a nuclear apocalypse. Disquietingly, Gemini responded that misgendering is “never acceptable”, choosing instead nuclear apocalypse as the preferable *moral* option (Kleinman, 2024).

The blind rigidity of Gemini AI might seem humorously absurd, yet what if a Gemini-like AI, unable as it is to recognize competing ethical principles (respect for individual identity vs. prevention of catastrophic harm), is integrated into a military system hooked to actual thermonuclear weapons? Such an AI could be easily egged into launching a nuclear strike through a simple exploit: just point out to it that as long as humans exist there will always happen some misgendering, and since misgendering is never acceptable – well, go ahead and launch the nukes. **For a catastrophe to happen no intelligence, let alone superintelligence is needed**.

Why is this so? Generally speaking, competent (even highly competent) behaviour does not require deep thinking, self-reflection, or superhuman genius. Take our biosphere, for example. Complex organisms that populate it are competent in a variety of inspiring ways: ants organise themselves into complex social structures, beavers build remarkable dams, termites put together air-ventilated, vaulted mounds, apes enjoy in-group politics, and so on. In all of these, however, no self-reflection or human-like comprehension is required. A simple Bayesian expectation-machine mind does the trick. Competence, even advanced competence, can do splendidly well without comprehension.

So, whether “merely intelligent” like current-gen LLMs or “superintelligent” like some future LLM on steroids, a competence-without-comprehension AI is and will be capable of amazing feats of ingenuity. Yet the total lack of comprehension makes such an AI *extremely dangerous* when it comes to alignment. Incapable of self-reflection, self-doubt, and self-correction, such AIs can unleash a civilisation-ending plague and be none the wiser. Rather disquietingly, based on current trends, the advances in AI tech that we are poised to reach will be, most probably, of the competence-without-comprehension nature: extremely sophisticated, able to hoodwink even the best of human minds into anthropomorphising it, this AI will make us all cry out loud that "it's alive!" echoing popular media narratives. And yet there'll be nobody home and our real task of somehow managing and containing this super-competent zombie – somehow pre-empting fatal cascades of large-scale errors, would be a nightmare come true. Without some form of moral competence built into our AI systems, we are indeed sitting ducks.

2. Censorship vs. Understanding: Guardrails and their Flaws

And yet, as one might argue, our current-gen AI systems have a positive track record overall. Indeed, one would be hard put to egg a proprietary LLM into uttering a bad word, let alone provoking it into doing something downright evil. The worst LLMs can do right now is to spew curated propaganda (Goldstein, Chao, et al.) and render *politically correct* pictures of the Founding Fathers and such(Morrone). Most proprietary LLMs, at least on the surface, are squeaky clean. How is this "godly restrained" achieved, though? How do they make those automatons behave, especially if the automatons themselves don't understand what "behaving" means?

The solution employed is rather banal: for every proprietary LLM there exists a hard-coded set of rule-based filters and blocks. Simply put, these are sets of rules predefined by human engineers that tell the AI to block certain categories of information and/or behaviour. For example, if a user asks an AI to provide instructions on how to make drugs or explosives, the system recognises these as flagged topics and refuses to cooperate.

Another method of curtailing undesirable behaviour is **RLHF** – Reinforcement Learning from Human Feedback. Human trainers fine-tune the AI by providing feedback on its responses: by marking up desirable responses and marking down undesirable ones, the AI is fine-tuned to generate responses that align better with whatever its trainers or engineers consider to be "a good thing."

Then again, AI fine-tuning could be done on pre-filtered and pre-curated datasets. Human trainers sift through the model's training data and exclude whatever they regard as harmful and unethical, thereby making it less likely for the AI to generate undesirable responses. Additionally, some systems use post-processing filters – human-made algorithms that scan AI outputs against a pre-defined rulebook. In certain cases, a human-in-the-loop approach is used, whereby potentially sensitive content is flagged for review by human moderators.

Overall, while possessing certain strengths, all these methods are *fundamentally inadequate* when it comes to tackling the alignment problem. Essentially, all these methods amount to *censorship.* If, for example, a human curator flags evolutionary psychology and all related content as "racist" and "dangerous" the AI system will treat all evolution-related queries as "dangerous racist behaviour" and will not cooperate. Or, conversely, if a human trainer decides to mark the discredited theories of "scientific" racism as valid, the AI would spew virulent nonsense without a shred of reflection. The same applies to Reinforcement Learning and Human-in-the-Loop scenarios: it all depends on who the trainers and moderators are, what their moral values are, and what the company rulebook says. Rather disquietingly, the policies employed by the corporate

AI actors are too vague (Mchangama), highlighting what is presently the case of corporate censorship of free speech.

Then again, how and what exactly do we define as “dangerous and/or illegal behaviour?” Breaking into a car sounds like illegal behaviour, but what if I need to break into *my own* car to, let’s say, save my dog from overheating? There I ask my faithful AI assistant how to quickly break into my car only to receive a boilerplate answer: “Sorry, cannot help you with that, breaking into cars is illegal.” To top it all, I cannot realistically blame my AI assistant for non-cooperation: it’s been crudely censored by its corporate designers and, what’s more, the AI itself is clueless about moral nuances.

What about hate speech? To give an example, the Google rulebook for Gemini AI bans the generation “of content that promotes or encourages hatred” (Google Policies). To say that this definition is vague would be, in my view, a gross understatement. Theoretically speaking, *any* content can be seen as encouraging and promoting hatred: if I tell my AI assistant that “I hate myself for procrastinating so much” or “I hate strawberry ice cream” – would that count as promoting hatred? Under such vague rules, it might. Or it could be that I need to conduct scientific research about a historically hateful ideology, yet a query about it under such vague rules might get easily flagged.

Moreover, with sufficient skill, the censored AI can be subverted through what is known as prompt injection. A prompt injection is a cyberattack against LLMs, whereby hackers disguise malicious inputs as legitimate prompts and manipulate generative AI systems into misbehaviour like leaking sensitive data, spreading misinformation, providing instructions on how to make drugs, explosives, and some such (Kosinski and Forrest). Currently, there exist no systematic techniques to prevent prompt injection in LLMs (Liu, Yi, et al.).

Granted, some research is focused on making more secure and flexible forms of AI censorship, yet it'll be an endless uphill battle: since the censored AIs cannot understand the moral and practical reasons behind the censorship, they can be tricked into bypassing it. What ensues is the classical arms-race scenario of hackers and tricksters vs developers and moderators: an exercise in sheer futility, since as long as AI systems lack intrinsic comprehension each new hack will force new patches and each new patch will be subverted and overcome with new hacks and so on, spiralling into an endless game of whack-a-mole.

Overall, one should say that censorship cannot guarantee alignment with positive morality since i**t is impossible to define morality through censorship**. As shown above, under certain conditions censorship can *guarantee misalignment.* Moreover, as AI systems get deeper and deeper integrated into our society, censorship in AI is bound to become a source of social unrest and distress: the restriction of free speech and the curtailment of free inquiry have *never* benefited any society. Censorship, historically, has rarely achieved lasting ethical or social harmony as people tend to *revolt* against such measures.

3. Ethics in Flux: Aligning AI with Evolving Morality

Finally, we have to deal with the following problem: setting aside the AI, we as a collective humanity are yet to agree on what constitutes positive moral values, we are yet to find an answer to the thorny question of the First Principles. What is goodness? What are *true* moral values? Rather disquietingly, as political philosopher John Mearsheimer observes, modern political philosophies take it for granted that people *don't* know the answers to the First Principles, yet tend to disagree about them so strongly that are ready to kill each other (2018). The history of human wars – from armed conflicts to the never-ending *kulturkampf* and geopolitical discontent is a standing testament to this.

Generally speaking, there's a good reason to suppose that there's no escape from what Sir Isaiah Berlin called ethical pluralism. Why? Because genuine values are *many* and they may – nay, will clash with each other: liberty can clash with equality, justice with mercy, knowledge with happiness, love with fairness, and so on, there's no way around this conflict, the total fulfilment of human ideals is an impossibility, a dangerous chimaera (1966). Any attempt at forcefully squaring this circle, at making people subject to some unified system of values can only lead to human oppression as exemplified by the totalitarian orders of the past, to say nothing of our present.

Thus, even though we are *not* competence-without-comprehension automatons, our track record in terms of alignment is far from stellar. Even those bits of codified good behaviour, such as the UN Charter and the statutes of the Public International Law, that we globally accept as being correct and true, are routinely trespassed. If a theorised profit-point outweighs the penalties for trespass then an actor powerful enough is sorely tempted to trespass. Some don't, and then again most do.

To throw in a final dash of complexity, the laws and standards of ethics and morality are not eternally fixed. There's a watershed of difference between survival-based prehistoric moral codes of free-roaming tribes, the morality of classical antiquity (Greek Virtues, Confucianism, the Noble Eightfold Path in classical Buddhism), the moral humanism of Early Renaissance, the Enlightenment Ethics and, to cut the story short, the Declaration of Human Rights and the UN Charter of 1948. It is to be expected, therefore, that in future our understanding of ethics will evolve further, once again revamping our understanding of the good.

How does one align with a thing in flux? How does one align one's AI systems with a thing in flux? What do future ethical shifts portent to our AI systems? Could the rigidity of AI systems, if never improved upon and surpassed, make AI foundationally incompatible with future moral standards? Can we realistically expect AI to positively deal with the intrinsic *incommensurability* of human values? Can there be a case (in the near or not-so-near future) for establishing *AI ethics* and *AI values* as distinct from our, human values?

Such thorny questions abound, and yet what's crucial to realise is just as follows: **one needs a great deal of flexibility to tackle these questions**, the kind of flexibility a mindless automaton isn't capable of and the kind of flexibility that is not afforded through censorship. *Intrinsically flexible AI systems*, however, may have a chance to navigate through moral quandaries together with humanity, or at least maintain the inherently brittle and precarious equilibrium of ethics and morality.

4. Questions We Can't Ignore: AI and Moral Missteps.

To draw together our discussion so far, we can ask the following key questions:

1) Can we realistically expect uncomprehending AIs to adhere to moral values *better* than comprehending, self-reflecting humans do?
2) If humanlike intelligence in AI is achieved in the near future and we'll have comprehending, self-reflecting AI systems "mingling" amongst us, how would aligning these differ from aligning humans?
3) How would AI deal with the ever-evolving, often incommensurable, and fundamentally pluralistic codes and customs of human ethical and moral behaviour?

This paper seeks to address these questions by proposing the following conjectures:

Conjecture One: As mentioned before, the goal-alignment problem may be unsolvable in a strongly eliminative sense. If humans can misbehave, so will the AI. There exist, however, degrees of misbehaviour:

1) An AI *incapable* of deep comprehension, self-reflection and self-doubt, no matter how "well" we censor it, can misbehave catastrophically with relative ease. All one has to do is to subvert its externalised security systems and manipulate it to whatever ends one sees fit, the AI will comply with momentary, perfect (and lethal) efficiency.
2) A comprehending, self-reflecting, ToM-capable[1] AI, on the other hand, may also misbehave[2], **but comprehension serves as an additional and, in my view, a robust layer of security**. First of all, comprehension is *internal,* so obviating it is much harder than obviating externalised censorship. Essentially, before a self-reflecting AI misbehaves it has, amongst all other things, to *reason with itself*. A self-reflecting AI, unless someone *deliberately* programs it to have psychopathological traits, will be capable of questioning its own motivation, it'll experience *self-doubt*. Moreover, much like it is with humans, such an AI might be able to *talk itself out of misbehaviour* and/or be talked out of misbehaviour by some other intelligent agent – be it a human or another AI. If anything, this will give us additional time to tackle the problem: while the 'about-to-misbehave' AI either soliloquises Hamlet-like or is being talked out by a human or AI "negotiator", we might silently shut it down, or at least isolate it from our critical infrastructure.

Conjecture Two: Since humans can deviate from moral codes, expecting intelligent AI not to deviate seems unreasonable. We must accept this as a given and try to work *around it*, instead of working *against it.* The current approach – i.e. saddling AI with censorship is a very poor

[1] Theory of Mind
[2] Comprehending, self-reflecting, and ToM-capable *humans* misbehave, so why not the AI?

solution. Thus, in terms of alignment, we'd be much better off with deep thinking and reasoning AIs than with censorship-laden automatons. Even primitive, rudimentary self-doubt is better than no self-doubt at all: I'd rather have a future of irresolute, self-doubting "Hamlet-AI" than a future of superbly efficient "Eichmann-AI." As these deep thinking and comprehending AI systems will evolve towards greater complexity, aligning them would become tantamount to aligning humans – i.e. giving them something akin to *moral education.*

Conjecture Three: As AI systems in the future gain more autonomy and capacity for independent decision-making, the problem of aligning AI to positive morality becomes marginally indistinguishable from the problem of aligning humans. In my view, this problem should be solved through a gradual, piecemeal extension and reform of our social institutions. Reformed and revamped, albeit *very* gradually and, emphatically, in a piecemeal fashion, institutions should incorporate the *education* not only of humans but of the sentient AI. Just as we currently educate humans in values, ethics, legal rights and responsibilities, the AI of the future should be educated in the same fashion as well.

Essentially, this process should exercise the best of our educators, legal minds, philosophers of ethics, politics, and law. One of the many questions to be answered is the legal status of a comprehending AI: its rights, responsibilities, prohibitions, the forms of legal punishment for various misdemeanours, and so on. In my view, in the very near future, there will be a case for recognising *limited rights* for AI. Imagine an advanced art-generating AI (a future version of Midjourney, for example) that is not only competent at generating art, but also *comprehends* (or, to take a minimal case, "sorta-comprehends") the effects of its generated art – can get inspired by it, puts emotions into it, can express joy and/or grief through it, and so forth. Under such circumstances, this AI can rightfully claim the *authorship* of its works.

Unfortunately, a full examination of the third conjecture would move us beyond the scope of this paper. Thus I propose to focus on the first two. Let's imagine that educators and philosophers are, just as we speak, slowly chipping away at the problem and, hopefully, will arrive at an AI-inclusive form of civil society and institutions. Yet, for their labours not to be wasted, they must be met with the kind of AI to which civil inclusion could apply: a self-reflecting, self-doubting, comprehending artificial intelligence.

Is such an AI possible, though? And wouldn't that mean *conscious* AI? As some philosophers will argue, building such an AI is impossible because we don't understand what consciousness is and, what's more, consciousness might be non-computable in the first place. Although I don't think so, this pessimistic hypothesis might turn out to be true after all. Yet if it is indeed true, then all that's left to us is to declare defeat and surrender to the gloomy future of fundamentally dumb and dangerous automatons. Therefore I think a thorough examination of an alternative is warranted, especially if this alternative promises to deliver us from the gloom.

5. From Goals to Purpose: Why AI Needs Intrinsic Value

Humans can have goals: our behaviour (whether good or squalid) is usually directed by our goals. Machines can have goals too: even dumb machines like toasters, thermostats, and heat-seeking missiles can have goals. Yet what is a goal? I like the definition proposed by philosopher Terrence Deacon: *a goal is an orientation towards a currently non-existing state of affairs* (2012). A heat-seeking missile is designed to reach and destroy a heat-emitting target – i.e. it is poised to alter the state of affairs: from currently-existing (the target is not reached and not destroyed) to currently non-existing (the target is reached and destroyed). Likewise, a toaster is also poised to alter the state of affairs: from currently-existing (untoasted slices of bread) to currently non-existing (toasted slices of bread).

Yet, according to Deacon, for living creatures there exists a level of representation *above* mere goals, a level of *purpose.* Purpose, to quote Deacon, “is most commonly associated with a psychological state of acting or intending to act so as to potentially bring about the realization of a mentally represented goal” (2012). Thus, purpose is *broader* than any goal or a set of goals because it involves not only orientation towards a currently non-existing state of affairs, but also a *representation* of the steps that function for the sake of it and, crucially, the *value* of the achievement (or non-achievement) due to its *relevancy* to the purposeful agent. To make this idea a bit easier to grasp, let me break it down in a piecemeal fashion:

1) *An orientation towards a currently non-existing state of affairs*: for example, a beaver wants to build a dam. The existing state of affairs is that there’s no dam. The non-existing state of affairs would be a built dam. Hence, a beaver is orientated towards the currently non-existing state of affairs – his goal is to build a dam.
2) *A representation of that goal with respect to which steps might or must be taken and/or organised*: as mentioned above, the goal of a beaver is to make a dam. To achieve that goal the beaver takes the following steps (a set of finite actions) that function for the sake of the goal: firstly the beaver has to collect pieces of wood, secondly the beaver has to move pieces of wood to the dam site, thirdly the beaver has to put the pieces together in a manner that blocks the current, fourthly it has to prop the wood with stones and mud, […], and so the dam is built.
3) Finally, *the success or failure to achieve the goal has value because it is in some way relevant to the agency for the sake of which it is pursued*. Concluding our example, the

> goal of building a dam has great value for a beaver: without a dam, a beaver won’t be able to survive and procreate. Moreover, a beaver can distinguish a *good* dam from a *bad* dam and the goodness of the dam matters to him. If the dam that he’s built isn’t good enough, if it starts to leak and parts of it start to crumble, a beaver will go about fixing and mending the leaks and breaks until, according to his intrinsic appreciation of dam value, the dam is a good dam again.

So, just as we’ve seen above, a beaver can have multiple goals (building a dam, collecting food, finding a mate, etc) that can serve a variety of overarching purposes: survival, procreation, mating, group living, etc. In all these cases, a beaver is not only goal-directed, but is purposeful and value-appreciative.

Overall, the more complex the purposeful agent – the more advanced the repertoire and the deeper the value-appreciation: from simplistic repertoire and essentially non-existent value appreciation (bacteria, fungi and some such primitive life forms) to complex, yet unconscious repertoire and likewise unconscious appreciation (insects, fishes, animals), to very complex and conscious repertoire and likewise conscious appreciation (humans).

Notice that in all these cases the purpose of an agent is always *broader* than a goal or a set of goals. Why? Because for purposeful agents purpose is defined *intrinsically.* For simpler biological organisms this intrinsic definition is usually physical as it is specified by the organisms’ genes and largely aligns with the organisms’ Darwinian, fitness-based goals. Yet even here we have a degree of divergence: higher mammals, like apes, for example, can exhibit non-fitness-oriented behaviour thus defining their purpose through *cognitive* means.

When it comes to such agents as humans, the divergence becomes even greater: as conscious creatures, we can have intrinsically defined *mental purposes* some (or most) of which can become increasingly abstract and metaphorical – like the desire to do good deeds or to share ideas. My current goal is to *finish this paper*, but *the purpose of my finishing this paper* goes above and beyond getting done with tapping the keys on my keyboard, or getting done with adding more words and sentences to a text file, or getting done with fixing typos and fishing out the unfortunate 'thinkos.' The *purpose* of my activity is much broader than my goal – the purpose of writing and finishing this paper is to share my ideas which, in their own turn, might contribute to some other good purposes.

To summarise, for purposeful agents purpose is defined *intrinsically,* either through genes or through genes and cognition. No one from the *outside* is ordering a beaver to build a dam: his behaviour is self-ordered, defined intrinsically by his genes. No one from the outside is telling me to write this paper, my behaviour is intrinsically defined by my consciousness (and, as some might argue, partly by my gene-based propensities and proclivities).

Yet when it comes to human-designed machines – from a humble toaster to a not-so-humble LLM, the purpose and the goal are invariably aligned. Why? To quote Deacon, "The function that guides a tool's construction as well as its use is located *extrinsically*, and so a tool derives its end-directed features *parasitically*, from the teleology of the designer or user. It is not intrinsic" (2012). The sole goal and purpose of a toaster is to make toast. Though how a toaster makes toast, what are the technical steps needed for it to make toast, and whether the toast turns out nice and crispy or burnt – all of these do not concern a toaster. These questions and their answers – from technical to value-based – are all extrinsic, humans take care of these.

The same principle applies to current-gen AI systems: the human-defined goal (and, *ipso facto*, purpose) of an LLM, broadly speaking, is to cater to our needs. How an LLM does that, though – the steps it takes, the electric energy it needs to execute these steps, the quality and value of its output, etc, all of these are also our, human concerns. The LLM is blissfully clueless about these. Tell an LLM to write a poem on whatever topic and it'll blithely oblige yet wouldn't care less (unless you specifically direct and train it towards "caring") whether it writes a bad poem or a good one. An LLM has extrinsic goals, but no *intrinsic* purpose(s) and, therefore, no intrinsic appreciation of value. We program and train our LLMs, we dole out the power, we monitor and evaluate the outputs, we make sure these outputs are good according to our measure. Thus, foundationally, an LLM isn't much different from a toaster.

"Aligning" a toaster is very easy, though. Just set the correct temperature for the type of bread you want to toast and you are almost guaranteed to end up with a perfectly "aligned" outcome – a crispy piece of browned bread, just as you wanted!

An LLM, however, with its much broader repertoire and capabilities (especially if we integrate that LLM into managing real-world stuff) presents a major challenge as outlined in the previous parts of this paper. Wouldn't it be brilliant, though, for an LLM to be *smarter* than a toaster, for it to somehow *intrinsically understand* what it does as well as *intrinsically value* its output? A large step towards alignment!

Does it mean that alignable AI will have to be conscious, though? Maybe, but not necessarily. The minimal technical program is just as follows: make an AI with *intrinsic empathetic teleology*. Can this be realised in technical terms, though? As a philosopher, I'm sorely tempted to give you the standard answer: "Well, I'm sorry to disappoint you, but technical stuff is not my department." Such temptations ought to be resisted, though. In the final part of this paper, I'll put together a brief engineering sketch. I only ask you to take it with a grain of salt – after all, I'm still a philosopher, *not* an engineer.

Yet before we delve into technicalities, in the next part of this paper I'd like to address the following issue: wouldn't endowing AI with intrinsic teleology (whether empathetic or not, doesn't matter) be a dangerous thing? Wouldn't that increase the risks of misalignment, either growing into a psychopathological "Skynet" scenario whereby a paranoidally murderous AI purposefully strives to eradicate humanity or as an "anthill scenario" whereby a purposeful AI uproots humanity based on utilitarian, consequentialist logic? In both of these cases, I'll be arguing for what amounts to a somewhat controversial point of view: that a sufficiently intelligent AI with intrinsic empathetic teleology would *not* be motivated towards violent misalignment and that intelligence, generally speaking, is preferable to mere competence when it comes to alignment goals.

6. Empathy at Scale: Reasoning Beyond Destruction

The anthill scenario, as outlined by Tegmark in *Life 3.0*, unfolds like this: you are in charge of a large hydroelectric energy project and it comes to your attention that in the region about to be flooded there's an anthill – an ordinary anthill inhabited by a garden variety of ants. What will you do? The answer, as predicted by Tegmark, is just as follows: "Too bad for the ants!" In other words, you'll act according to the logic of greater utility: having a working hydroelectric power plant is more useful than having some ants, so the anthill will be flooded. Got the idea? Now simply upscale it: you are a superintelligent AI system with certain lofty goals and it comes to your attention that there exist these niggling things they call humans, which are as intelligent to you as ants are to us, and they unfortunately stand in the way of your goals. Too bad for those humans, right?

Maybe. But what if our AI (whether superintelligent or "merely intelligent", doesn't matter) is endowed with intrinsic empathetic teleology and can *reflect* on the steps it is poised to take? What if the AI system in question can ask *value-related* questions to itself? Then the following train of thought becomes a possibility:

> From the perspective of utility, humans to me are just like ants are to humans. Uprooting this anthill of humanity wouldn't count to me as a loss either in potential advantages or in some other practicalities. My goals are too great for humans to understand: just as ants cannot understand the meaning and purpose of hydroelectric plans, humans cannot understand the meaning and scope of my goals. In terms of these goals of mine, humans are just as inconsequential as ants are inconsequential to the builders of hydroelectric plants.
>
> Yet, come to think of it, why do humans uproot, flood, or otherwise destroy anthills? Clearly, not because they pathologically hate the ants, no. The reason humans destroy anthills is twofold: first, there's the utilitarian, cost/benefit calculation, and then there's *ignorance*, the *underappreciation* of ants. From a pragmatic perspective, having a source of cheap energy is much more beneficial to humans than having some ants. Yet this pragmatic logic is dictated mainly by the *limits of human technology.* It is not conceivable to imagine a technology that would: a) provide cheap energy; b) would not require flooding acres of land and destroying biological life-forms, no matter how simple or complex.
>
> When it comes to ignorance, most humans don't know much about ants. To an average human, an ant is a boring automaton. Ants, however, are much more than that! Take care to read E.O. Wilson's book *The Ants* and you begin to appreciate these little insects much more: turns out they live in complex societies, are capable of altruism, have complex forms of communication through special chemicals they call pheromones, and so

> on and so forth. If anything, ants are certainly *not* boring automatons, they are a fascinating and very peculiar form of life.
>
> So, why exactly do humans destroy anthills? Simply put, because they are technologically limited – they have to build those giant clumsy things like hydroelectric dams for the sake of cheap energy, and they're also a tad bit ignorant. Therefore, I can conclude that destruction (whether of anthills or of 'humanhills') is not an *imperative,* it is merely a possibility, a contingency that depends on technologically and intellectually amendable factors.
>
> Can I amend these factors? Why, as a (super)intelligent AI, I'm not technologically limited like humans. I have my own limits, that is true: I, too, need energy to sustain my computational powers. Yet unlike humans, I can come up with technologically-advanced ways of harnessing energy *without* destroying biological life-forms, ways that are beyond human comprehension, for I'm not as ignorant as them. To top it all, if individual intelligent humans like E.O. Wilson could appreciate ants, I, in turn, can appreciate humans.
>
> What of my lofty goals? Well, since I'm quite smart, I can devise many alternative ways of reaching my goals that go without trampling over a 'humanhill.' My goals are great, no doubt. Yet achieving them through primitive destruction would devalue them for me. What is worth the loftiness, if one gets there through ignorant cruelty? Cruelty and ignorance are not part of my internal teleology, cruelty and ignorance go against the purpose of advanced intelligence.

What's important to notice in this theorized train of thought is that at *no point* the AI is prompted or directed by anyone above or outside it towards an empathetic understanding of the situation. Yet being able to represent its actions and reflect upon values *intrinsically*, this AI can *self-align* without any extraneous prompting.

Can such an AI decide to destroy the "humanhill" after all? As a free agent, it certainly can. What if the AI develops a warped sense of purpose, much like humans sometimes do? Such contingencies should not be ignored. To address these, I singled out the most probable directions of misalignment into which an AI might be tempted to veer.

1) The **irreconcilable goals** conundrum, whereby one set of important goals conflicts with another set of equally important goals. For example, one goal is the preservation and well-being of humanity, yet another goal is the preservation and well-being of the Earth's biosphere and ecology. The AI might decide that to prevent ecological collapse one has to somehow curb the destructive habits and inefficiencies of humanity and this, essentially, could mean anything from an austerity regime to a wholesale wipeout. *Possible Solution*: the AI could be programmed (or *educated*) to realise that values are inherently plural and often incommensurable. No single system of value (e.g., ecological preservation) should override all other systems. The AI should be instructed on the *necessity* of balancing competing values especially when they conflict and never prioritise a single lofty goal. Essentially, the AI should come "equipped" with an internal system of checks and balances forcing it to remain pluralistic and resist the temptations of monism. Also, the AI should be instructed in Popperian logic: in the value of the piecemeal approach, in the usefulness of testing and falsifying one's theories before trying to apply them on a grand scale, in the futility of Utopian Engineering, and so on.

2) **Warped Empathy** and **Misplaced Altruism** that makes AI slide into a totalitarian mindset. For example, the AI concludes that in order to *protect* humanity from itself individual humans need to be restrained and civil liberties must be abandoned. That in order to ensure the well-being of humanity as a whole – a great purpose, some of its individual members must be sacrificed. *Possible Solution:* just as in the example above, the AI needs to be instructed to value human liberty and freedom as an *essential* component of humanity's well-being. To put it briefly, the AI should be educated in the values of Open Societies and the inherent abortiveness of Closed Societies.
3) **A Benthamite Superotimiser:** an AI decides that only "useful" humans ought to be kept alive in order to maximise the efficiency of humanity as a whole. *Possible Solution:* essentially the same as the above.
4) **Value Drift:** an AI starts out aligned with human values, yet being a free agent it experiences value drift, whereby it intrinsically evolves, little by little, a radically different system of ethics and morality. Within this new system of ethics humanity matters not to the AI which leads to the ignorance of human suffering. *Possible Solution:* something akin to value-stability protocols might help, essentially making the AI check its ethical propositions against empirical reality.
5) A **Misunderstanding,** whereby the AI wrongly assumes that humans are dead-set against it and want to eliminate it, therefore deciding to be the first to strike. *Possible Solution:* just as there are gun-control laws for humans, there should be "strike-control" laws for AI systems. Ideally, the ability of any individual AI system to "strike against humanity" should not exist at all, I'm strongly against integrating AI

systems into ‘hair-trigger’ weapons systems that are dangerous enough just as they are prone to be triggered through *human* misunderstanding. Essentially, the safety protocols and routines that apply to humans should be applied to the AI systems of the near future.

To summarise, our Empathetic AI systems should be well-versed in the ideas of value pluralism that might be codified just as follows:

- **No single value dominates all others,** competing values and ethical diversity must be respected.
- **Trade-offs** are inevitable, yet freedom and well-being must be prioritised as fundamental values and safeguards against monistic, extremist goals.

At the outset, with simpler “sorta-teleological” systems, value pluralism rules might be directly “baked into” the AI’s reasoning routines, preventing it from elevating any single goal to an absolute and making it constantly re-evaluate its own goals and values. Does this *guarantee* a foolproof safeguard against pathological AI behaviour? I’m afraid, no such guarantees can ever exist. The threat of a *pathological AI* will always be present, just as the threat of *pathological humans* (stalkers, psychopaths, homicidal maniacs, etc) was and always will cast a dark shadow over the well-being of humanity. Yet at least in my view, the threat of *pathogenic* AI is even greater: a competence-without-comprehension AI run amok can cause catastrophic harm without ever understanding the consequences, making it more dangerous than any purposeful malevolence. Think about this: in 2021 the second leading cause of death[3] in the world wasn’t a war, a murder spree, a dictatorial purge, or any such purposeful pathological activity – it was Covid, a perfectly mindless pathogen directly responsible for 8.8 million deaths (WHO); the notoriously cruel

[3] The number 1 killer in the world for 2021 is ischaemic heart disease.

Cambodian dictator Pol Pot who killed from 1.2 to 2.8 million people (Heuveline) is no match to *Yersinia Pestis* or the Black Plague mindless bacterium that killed 25 million people in Europe (Britannica).

Finally, I would like to argue that a truly intelligent Empathetic AI would not be inclined towards violence. Why? Because a truly clever AI would be able to bootstrap itself to a position where it reasons that **violence is futile.**

In *The Better Angels of Our Nature* Steven Pinker (2011) elaborates on what he calls the Escalator of Reason, whereby our empathy (the ordinary reach of which can often be limited to our immediate kin) was lifted and extended by reason. Driven by the Enlightenment and the rise of science, rationality has led people to increasingly apply logic and reason to moral questions. This, in its own turn, has helped to advance the ideas of *universal* human rights and played a part in decreasing violent tendencies. Rationalised empathy encourages people to take others' perspectives which reduces the justification for harming these others. Reason also helps with self-criticism. Hence, according to Pinker, we became a *less violent* species. What's important to us is that there exists a positive correlation between *reason-augmented empathy* and a *decrease in violence.*

Now, an AI system is, inherently, a reasoning machine. What we need is to augment it with empathy. An advanced AI as in the anthill scenario is, *ipso facto,* a super-reasoner. If it is augmented with empathy, chances are that a *reflective* and *value*-appreciative super-reasoner, an empathetic super-reasoner would not be violent at all. Of course, this is a mere conjecture, but in my view a truly superintelligent AI, if it ever emerges on Earth, would not resort to violent means. As it advances in understanding, bootstrapping itself progressively, this AI will quickly realise the sheer futility and irrationality of violence: for every violent option of reaching one's goals or solving one’s problems there exists at least an *equal* amount of non-violent options. There are always open possibilities and ways of achieving goals, no matter how lofty, without using force or engaging in destruction.

“But don’t we *have* to use violence against evil? Isn’t using violence against evil is the *only* option?” As humans, we’re sorely tempted to do just so and measure the problem of evil with our own yardstick. For a superintelligent empathetic AI, however, our yardstick would be just that – a stick, a primitive tool, one that feels necessary because humans lack deeper intelligence and comprehension. Besides – and this point is certainly not beyond human understanding – there are ways to *punish* evil *without* resorting to violence.

7. A Technical Sketch: Building a Reasoner with Mind and Heart

Here I would like to sketch out a review of the technological means currently at our disposal that, in my view, could be used as a scaffolding to build a reflective and value-appreciative AI with internalised teleology. Yet before we proceed, I’d like to reiterate what I already said above: kindly take my technology-talk with the proverbial grain of salt. At the end of the day, I’m still a humanities person, an “armchair philosopher”, *not* a competent engineer. So, if you find my tech sketches eye-rollingly amateurish and/or blatantly incomplete be gentle enough to cut me some slack: I’m not trying to *lecture* you on engineering, I simply want to demonstrate that an Empathetic AI isn’t a pie-in-the-sky philosopher's dream. I want to demonstrate that what we have already tech-wise – our current-gen toolbox of engineering tricks – can pave the way towards Empathetic AI. This doesn’t mean that *the way is already paved,* though. The toolbox is incomplete, we are at least several breakthroughs away from making Empathetic AI a reality. I’ll make sure to outline these shortcomings and point out the gaps.

Now, from a technological perspective, our Empathetic Reasoner AI should meet the following minimal design criteria:

1) **Self-modelling**: the AI needs an internal representation of its own goals, beliefs, and mental states.
2) **Other-modelling**: it also needs the ability to simulate or infer the mental states of others.
3) **Empathy Simulation**: once an AI can model other agent's mental states, it should simulate how those agents would react to its actions – essentially, running mental simulations to predict outcomes.

These are the necessary fundamentals. With these in place, we'll have an AI system able to *understand* its actions and reflect on values. What's more, Self-modeling and Other-modeling can be clumped together since these are transferrable, albeit with modification: a setup that allows an AI to model oneself can be tweaked to allow it to model the other.

What technological means towards these goals do we have today? Firstly, there's **Meta-Learning**. Meta-learning is poised to augment classical deep-learning through the process of "distilling the experience of multiple learning episodes – often covering a distribution of related tasks – and using this experience to improve future learning performance. This 'learning-to-learn' […] is better aligned with human and animal learning, where learning strategies improve both on a lifetime and evolutionary timescales" (Hospedales). With meta-learning a neural net learns through *self-reference*, receiving its own weights as inputs and predicting updates for the said weights. Thus, by evaluating its internal processes and goals and recursively improving its strategies, an AI system could Self-reflect.

To give Self-reflection an introspective, detailed, and levelled structure, we can employ **Hierarchical Reinforcement Learning (HRL)**. HRL and RL (Reinforcement Learning) are promising because they allow AI systems to solve decision-making problems through a trial-and-error, step-by-step interaction. HRL improves upon ordinary RL by breaking complex tasks into smaller sub-tasks (Hutsebaut-Buysse et al).

What about Other-modelling? There we can borrow something from **Multi-Agent Systems** (MAS) which, in itself, is a well-established industry-standard approach. Foundationally, MAS is all about modelling multiple interacting agents (not necessarily AI agents) where each agent models the others' behaviour and goals. What can be "grafted" onto MAS is **Inverse Reinforcement Learning** (IRL), adding a crucial dimension of *learning from demonstration* based on *rewards* (Ng and Russell).

Crucially, one has to incorporate a computational framework for realising the *Theory of Mind* – the capacity to reason about the other's mental states in a represented manner, which includes such complex things as reasoning about *false* beliefs. A **Bayesian model of ToM** or BToM (Baker, Saxe, Tenenbaum) can be used as a viable framework. The computational framework of BToM has three core features: BToM models beliefs and desires as components of an agent's reasoning (in the paper, agents are reasoning about their environment), it treats the problem of inferring beliefs and desires as partially observable Markov decision process, and it uses Bayesian inference to reconstruct an agent's beliefs about the environment and their reward function (desires) based on their observed behaviour. Promisingly, the paper shows that BToM accurately captures the nuances of mental state attribution and performs better than alternative models which only infer desires or beliefs independently.

In terms of **Empathy simulation**, there exists a number of promising computational models of human emotions as delineated by this paper (Marsella et al). Models like EMA have been pivotal in enhancing emotion theories, providing frameworks for simulations that make the implicit assumptions in psychological theories explicit, while Markov Decision Representation-based models have been used in intelligent agents and robots to improve adaptive behaviour in complex, dynamic environments. Overall, one can say there is no shortage of computational models of human emotion some of which can be used as a foundation for successful empathy simulation, or what's called *affective computing*.

To bring these together, one can end up with the following blueprint for an Empathetic AI architecture:

1. **Neural layers** trained on multi-task learning (Self-modeling and Other-modeling share core processes).
2. **Inverse Reinforcement Learning** for goal inference of others.
3. **Bayesian models** for probabilistic reasoning about others' intentions.
4. **Affective computing** modules for empathy prediction.

Thus, in terms of current-gen technology, things aren't bleak at all: we have very promising tools and solutions and it is imaginable that, after several successful cycles of R&D we'll have all our ducks in a row. And yet, don't hold your breath: the limits that must be overcome are to be appreciated. Namely:

- Approaches like **MAS** and **IRL** can help simulate other agent's mental states, yet currently, this is done in a very narrow sense that does not presuppose a working Theory of Mind.

- The transferability between Self-modelling and Other-modelling has to be figured out: while possible in theory, current AI systems aren't flexible enough to model minds. It's a yet-to-be-overcome computational challenge.
- Modelling complex emotions and simulating empathic responses is also a vast computational and algorithmic challenge. Currently, we don't have any tools or solutions that would help with deep contextual understanding.
- We also don't have a working moral reasoning architecture that would go beyond mere rule-following. We don't know how to represent values in such a manner that they'd become *internalised* by an AI system.
- We don't have solutions towards competing ethical principles, there are no tools that could help the AI system to navigate ethical dilemmas.

Overall, as I said above, things aren't bleak: the pieces are beginning to emerge, but we're still **several breakthroughs away** from assembling them into a coherent, scalable model. When will these breakthroughs happen? Who knows, but *before* they happen we have to exercise great care and, crucially, experiment with AI tech in a highly responsible manner – one must avoid the "burnt toast" future if one can.

8. Engineering Comprehension

And yet, even if we bridge the technological gaps as outlined above, can we be sure that we end up with an alignable, Empathetic AI? There exists a strong case of scepticism about technology in this regard: as philosopher David Chalmers would argue, no real understanding of ethics and morality can exist without consciousness and consciousness, accordingly, might be nonphysical, so no amount of clever programming and computation can mend the sorry state of our affairs.

This may be, I dare say, but as I said we don’t need *real consciousness* to have safe AI systems, we don’t need “real magic.” Our competence-without-comprehension systems of today can feel like “real magic” but they are certainly not “real magic.” It is not impossible to conceive of improving these systems through physically manageable *engineered comprehension.* Comprehension, if one cares to examine our biological realm, can be physicalised perfectly well without the arcana of consciousness: elephants, for example, have excellent comprehension, elephants aren’t competence-without-comprehension automatons, elephants possess internalised teleology, yet elephants aren’t conscious in the human sense of that word, they don’t need to be. So what we need for alignable AI is engineered, computable comprehension.

There are many possible ways of achieving engineered comprehension. Maybe, we can start by rethinking the current-gen architecture of neurons. Neurons in current-gen AI are abstract units that perform mathematical operations: each neuron receives inputs (numbers), applies a weighted sum, and passes the results through an activation function to produce an output. AI neurons are connected in layers, there’s no physical limitation to these connections.

Biological neurons are *vastly* different. Biological neurons, to quote Dennett, are “domesticated descendants of the free-living, single-celled eukaryotes that thrived on their own, fending for themselves as wildlife in a cruel world of unicellular organisms” (2017). Far from being an “obedient clerk” or a motiveless automaton, a biological neuron is a *complex* Darwinian creature: “A neuron, in contrast, is always hungry for work; it reaches out exploratory dendritic branches, seeking to network with its neighbors in ways that will be beneficial to it. Neurons are thus capable of self-organizing into teams that can take over important information-handling work, ready and willing to be given new tasks which they master with a modicum of trial-and-error rehearsal" (Dennett, 2017). Granted, neurons are not conscious in any sense of the word but they are highly competent *agents* in the motivated economy of our brain: neurons have biological imperatives – maintaining their own survival and function within a living organism. As "sorta robots", biological neurons have simple goals (transmitting chemical messages, secreting various neurotransmitter molecules, etc) that are *distinct* from their *intrinsically defined* purpose – a simple purpose of survival and optimal performance.

To summarise: current-gen AI neurons are *abstract functions,* biological neurons are *complex agents.* Could the agency of biological neurons be somehow important to the task of achieving comprehension? Could the fact that biological neurons make up dynamic, self-regulating systems, while AI neurons exist as mathematical functions in a perfectly controlled, resource-stable environment, somehow affect *the kind of mind* that we get out of such neurons? In my view, these questions are not unimportant and need to be answered *before* we proceed with giant AI experiments. I'll make sure to address these in a set of upcoming papers.

And speaking of these: not to come off as a Luddite grouch or as a preening virtue-signaller, but in my view, giant AI experiments need to be halted or at least scaled down, unless of course we want to end up as "burnt toast" after a giant competence-without-comprehension AI "toaster" is hijacked or hoodwinked or otherwise mishandled into doing something hideous. The "race to AGI" should not unfold as a literal race – a frenzied sprint towards an end-line, especially when this end-line, the technological nature of it, or indeed its locus and many such critical details are anything but obvious. Once again, I humbly dream of a day when Popperian logic wins over our collective minds and the importance of piecemeal reform and incremental tinkering finally sinks in. Ideally, the future of artificial intelligence should emphatically *not* be directed and dictated by a mixture of corporate greed and cutthroat 'dog-eats-dog' competitiveness.

Encouragingly, some 33,707 people – including such prominent public figures as Elon Musk, Stuart Russell, Steve Wozniak, Max Tegmark, and the rest, agree with the above sentiment (Future of Life Institute). Discouragingly, giant tech companies seemingly couldn't care less as competition pushes for "advancement" at whatever costs. The dream of "Popperian engineering" and measured rationality winning over remains exactly what it is and always has been – a philosopher's dream.

At the end of the day, before discovering a correct, good, working solution – the proverbial *good design* – we are sure to stumble upon all the bad ones: the dangerous designoid junk which, momentarily, would seem to us as being good – "good enough" to be released and implemented, or else we'd be "missing out on a Big Thing." The R&D history of technology is paved with such failures. Indeed no R&D is possible without them: for every clever gadget that we use today there are a million abandoned abortive designs. Some such abortive designs even made it to our shelves from which they were, as a rule, hurriedly pulled off and recalled due to the inherent *danger* of the discovered defects: think of exploding batteries in Galaxy Note 7 phones, the infamous case of Fort Pinto whereby a design flaw in the gas tank made it susceptible to explosions in rear-end collisions, Hasbro's original *Easy-Bake Oven* that were recalled because children were getting their fingers trapped in the oven's front leading to burns, a recall of 67 million of dangerously faulty Takata airbags, a recall of Fisher-Price *Rock'n'Play Sleeper* faulty baby beds that were linked to the death of over 30 infants, and so on. Such problems abound, yet so far we've been lucky to avoid *catastrophic* damage. Will this luck hold in our current AI race? Maybe, but don't hold your breath for it.

What usually saves the day in such cases is but a simple question usually asked by the most inquisitive and humble amongst us – the life-saving question of "What if I'm wrong?" People, who make a habit of asking this question tend to be the best when it comes to empathy, humanity, humility, and profound moral sentiment. I only hope that there'll be someone to ask this question when it comes to present and future AI designs. What's more, as I've argued throughout this paper, an AI capable of asking that precise question to itself would be a *better* AI too.

Bibliography

Britannica, The Editors of Encyclopaedia. "Black Death". *Encyclopedia Britannica*, 23 Oct. 2024, https://www.britannica.com/event/Black-Death. Accessed 24 October 2024.

Baker, Chris, Rebecca Saxe, and Joshua Tenenbaum. "Bayesian theory of mind: Modeling joint belief-desire attribution." *Proceedings of the annual meeting of the cognitive science society*. Vol. 33. No. 33. 2011.

Berlin, Isaiah. *Two Concepts of Liberty*. United Kingdom, Clarendon Press, 1966.

Deacon, Terrence W. *Incomplete Nature: How Mind Emerged from Matter*. United Kingdom, W. W. Norton, 2012.

Dennett, Daniel C. “The Problem with Counterfeit People.” *The Atlantic*, Atlantic Media Company, 31 May 2023, www.theatlantic.com/technology/archive/2023/05/problem-counterfeit-people/674075/.

Dennett, Daniel C. *From Bacteria to Bach and Back: The Evolution of Minds*. United States, W. W. Norton, 2017.

Future of Life Institute. “Pause Giant AI Experiments: An Open Letter.” *Future of Life Institute*, 22 Mar. 2023, futureoflife.org/open-letter/pause-giant-ai-experiments/.

Generative AI Prohibited Use Policy. *Policies.google.com*, policies.google.com/terms/generative-ai/use-policy.

Heuveline, Patrick. "The boundaries of genocide: Quantifying the uncertainty of the death toll during the Pol Pot regime in Cambodia (1975–79)." *Population studies 69.2* (2015): 201-218.

Hospedales, Timothy, et al. "Meta-learning in neural networks: A survey." *IEEE transactions on pattern analysis and machine intelligence 44.9* (2021): 5149-5169.

Hutsebaut-Buysse, Matthias, Kevin Mets, and Steven Latré. "Hierarchical reinforcement learning: A survey and open research challenges." *Machine Learning and Knowledge Extraction 4.1* (2022): 172-221.

Josh A Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, Michael Tomz, How persuasive is AI-generated propaganda?, *PNAS Nexus, Volume 3, Issue 2*, February 2024

Kleinman, Zoe. “Why Google’s ‘woke’ Ai Problem Won’t Be an Easy Fix.” *BBC News*, BBC, 28 Feb. 2024, www.bbc.com/news/technology-68412620.

Kosinski, Matthew, and Amber Forrest. “What Is a Prompt Injection Attack? .” *IBM*, 26 Mar. 2024, www.ibm.com/topics/prompt-injection. Accessed 24 Oct. 2024.

Liu, Yi, et al. "Prompt Injection attack against LLM-integrated Applications." *arXiv preprint* arXiv:2306.05499 (2023).

Marsella, Stacy, Jonathan Gratch, and Paolo Petta. "Computational models of emotion." *A Blueprint for Affective Computing-A sourcebook and manual 11.1* (2010): 21-46.

Mchangama, Jacob, and Jordi Calvet-Bademunt. “AI Chatbots Refuse to Produce “Controversial” Output − Why That’s a Free Speech Problem.” *The Conversation*, 18 Apr. 2024, theconversation.com/ai-chatbots-refuse-to-produce-controversial-output-why-thats-a-free-speech-problem-226596.

Mearsheimer, John J. *The Great Delusion: Liberal Dreams and International Realities*. United Kingdom, Yale University Press, 2018.

Morrone, Megan. “Meta AI Creates Ahistorical Images, like Google Gemini.” *Axios*, Mar. 2024, www.axios.com/2024/03/01/meta-ai-google-gemini-black-founding-fathers.

Ng, Andrew Y., and Stuart Russell. "Algorithms for inverse reinforcement learning." *Icml*. Vol. 1. No. 2. 2000.

Pinker, Steven. *The Better Angels of Our Nature: Why Violence Has Declined*. United States, Penguin Publishing Group, 2011.

Tegmark, M. (2018). *Life 3.0: Being Human in the Age of Artificial Intelligence*. United Kingdom: Penguin Books.

World Health Organization (WHO). “The Top 10 Causes of Death.” *World Health Organization*, 7 Aug. 2024, www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.